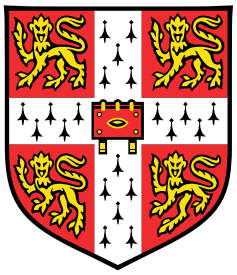
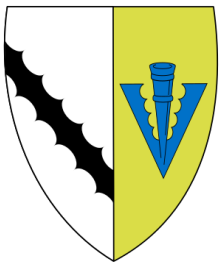


The Automatic Detection of Small Molecule Binding Hotspots on Proteins

Applying Hotspots to Structure-Based Drug Design



Christopher John Radoux

Department of Biochemistry
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Sidney Sussex College April 2018

For Emma

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Christopher John Radoux
April 2018

Acknowledgements

PhD students can face a wide variety of experiences throughout their project, both in terms of their research and their life around it. One of the most important factors is the support of the people around you, and in this regard I count myself extremely lucky. I have truly enjoyed my PhD project thanks to the support of my supervisors, colleagues, family and friends.

Firstly, I would like to thank my university supervisor, Professor Sir Tom Blundell. Before starting my PhD, I was told that having a well-known and important supervisor meant they would be constantly busy and not have time for you. Despite being exceptionally busy, Tom has an amazing ability to make time for his students, regardless of where he has just flown from or where he needs to be later that day. I am truly grateful for all of your support, it has been incredible to have the opportunity to work with such an inspiring scientist.

I would like to thank Tjelvar Olsson, Colin Groom, Beth Thomas and Alicia Higueruelo, who each supervised me at the CCDC. Tjelvar was the perfect supervisor for the first year of my PhD. He taught me *how* to be a researcher, a scientific programmer and a scientific writer. For the rest of my PhD, I often found myself asking "Would this result convince Tjelvar?"

Colin took over from Tjelvar as my supervisor in my second year, and I am extremely grateful for all of his suggestions and guidence during that time. He has an incredible imagination when it comes to asking insightful scientific questions, often providing three weeks worth work in our weekly meetings for which I still want to get through.

Beth took over from Colin during my third year, and guided me in turning a prototype method into something ready to be used by the scientific community, even coming up with the name, Fragment Hotspot Maps.

Alicia was my final CCDC supervisor, and took over as I was approaching the end of my project. Alicia is an amazing person, and I don't think anyone, who has had the chance to work with her, will ever forget her. As the time pressure started to build at the end of my project, Alicia did everything she could to support me.

I would like to thank my industrial supervisor, Will Pitt. Will has been a fantastic supervisor and mentor, providing key insight throughout the project. He ensured that I had a comfortable and productive time during my placement at UCB. Many of the ideas described within this thesis started life as "It would be cool if we could do…" in a discussion with Will.

I was the first PhD student to be based full time at the CCDC, and this has had a hugely positive impact on my PhD experience. I would like to thank everyone at the CCDC for being great to work with and providing an excellent support base. In particular I would like to thank Richard Sykes for the addition of the Grid API to the CSD Python API, which allowed me to do far more with the Fragment Hotspot Maps. I would also like to thank Pete Curran, who is now starting his second year of his PhD at the CCDC, it has been a pleasure working together in my final year and I wish him luck for the rest of his PhD (countless hours of python, proteins, endless reading …).

I would like to thank UCB and the BBSRC for funding my CASE studentship.

Thank you to Sidney Sussex college, and in particular Sidney Sussex Boat Club, for really enriching my time in Cambridge. It was wonderful to be a part of such a great community, and it has really defined my time in Cambridge.

I would like to thank my three closest friends, Krishan Sapra, Mark Dobson, Hansley Kalicharan-Goder. These guys continue to provide the perfect balance of support, fun, motivation and mockery to help me get through writing this thesis, and will be pleased that I have finally stopped being a student.

My family have always encouraged and nurtured the scientist within me, and have helped me through tough times to get me to where I am today. I am immensely grateful to my parents, Emma and Peter, my brothers, Joe and Nick, and to all of my extended family, who have shown interest in what I do, even when I can't quite explain it all.

Finally, I would like to thank my wonderful partner, Emma. You have been incredibly supportive while I have been writing, I hope I can repay you when you begin your own thesis. Of all the many reason why I have enjoyed my time here in Cambridge, being able to live with you (and Scrambles) is number one.

Abstract

Locating a ligand-binding site is an important first step in structure-guided drug discovery, but current methods typically assess the pocket as a whole, doing little to suggest which regions and interactions are the most important for binding. This thesis introduces Fragment Hotspot Maps, a grid-based method that samples atomic propensities derived from interactions in the Cambridge Structural Database (CSD) with simple molecular probes. These maps specifically highlight fragment-binding sites and their corresponding pharmacophores, offering more precision over other binding site prediction methods.

The method is validated by scoring the positions of 21 fragment and lead pairs. Fragment atoms are found in the highest scoring parts of the map corresponding to their atom type, with a median percentage rank of 98%. This is reduced to 72% for lead atoms, showing that the method can differentiate between the hotspots, and the warm spots later used during fragment elaboration.

For ligand-bound structures, they provide an intuitive visual guide within the binding site, directing medicinal chemists where to grow the molecule and alerting them to suboptimal interactions within the original hit. These calculations are easily accessible through a simple to use web application, which only requires an input PDB structure or code.

High scoring specific interactions predicted by the Fragment Hotspot Maps can be used to guide existing computer aided drug discovery methods. The Hotspots Python API has been created to allow these work flows to be executed programmatically through a single Python script. Two of the functions use scores from the Fragment Hotspot Maps to guide virtual screening methods, docking and field-based ligand screening. Docking virtual screening performance is improved by using a constraint selected from the highest scoring polar interaction. The field-based ligand screener uses modified versions of the Fragment Hotspot Maps directly to predict and score the binding pose. This workflow gave comparable results to docking, and for one target, Glucocorticoid receptor (GCR), showed much better results, highlighting its potential as an orthogonal approach.

Fragment Hotspot Maps can be used at multiple stages of the drug discovery process, and research into these applications is ongoing. Their utility in the following areas are currently being explored: to assess ligandability for both individual structures and across proteomes, to aid in library design, to assess pockets throughout a molecular dynamics trajectory, to prioritise crystallographic fragment hits and to guide hit-to-lead development.

Table of contents

List of figures

List of tables

Acronyms

K_d dissociation constant. 137

5-HT$_2$ 5-hydroxytryptamine type 2 receptor. 3

ADMET absorpotion, delivery, metabolism, excretion and toxicity. 1, 3, 16

API Application Programming Interface. 71, 75, 76, 79, 81

AUC area under curve. 103, 108, 110, 112, 114, 119, 122, 126, 127

CADD computer-aided drug design. 13, 71, 84, 92

CCDC Cambridge Crystallographic Data Centre. 25

CSD Cambridge Structural Database. 25, 27, 28, 40, 89, 99, 102

CSS Cascading Style Sheets. 74

D$_2$ dopamine type 2 receptor. 3

DMSO dimethyl sulfoxide. 28

DUD-e directory of useful decoys - enhanced. 104, 119, 142

EBI European Bioinformatics Institute. 129, 134

EF enrichment factor. 103, 108, 110, 112, 119, 122, 126, 127

FBDD fragment-based drug design. 8, 14, 17, 33

FFT fast fourier fransform. 39

FIM Full Interaction Maps. xvi, 89, 91

FTP File Transfer Protocol. 75

GA genetic algorithm. 96, 98

GE group efficiency. 29, 34, 36, 60, 64, 65

GUI Graphical User Interface. 82

HCS high concentration biochemical screen. 19

HEPES 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid. 64

HS-LS hotspot-based ligand screener. 106, 110, 111, 118, 119, 122, 127

HTML HyperText Mark-up Language. 74

HTS high throughput screening. 8, 14, 15, 17, 24

IMPDH inosine monophosphate dehydrogenase. xv, 27

IP intellectual property. 15

ITC isothermal titration calorimetry. 18

JS JavaScript. 74, 79

MCSS multiple copy simultaneous search. 23, 39

MD molecular dynamics. 32, 33, 39, 69, 93, 134

MMP matched molecular pair. 34, 36

MS mass spectrometry. 19

MSCS multiple solvent crystal structure. 22, 23, 46

MST microscale thermophoresis. 19

NMR nuclear magnetic resonance. 18, 21

nPDB-LS novel PDB overlay ligand screener. 107, 108, 110, 111, 118, 119, 122, 127

PDB protein data bank. 10, 73, 133

PDB-LS PDB overlay ligand screener. 107, 108, 110, 111, 118, 119, 122, 127

PKB protein kinase B. 36

PPI protein-protein interaction. 8

RCT randomised controlled trial. 5

RDA reporter-displacement assay. 19

ROC receiver operating characteristic. 103, 112, 119

SACP simulated annealing of chemical potential. 32, 33

SAH S-adenosyl-L-homocysteine. xvi, 98, 99

SAR structure-activity relationship. 1, 4

SMARTS smiles arbitrary target specification. 102

SMILES simplified molecular line input entry system. 13

SPR surface plasmon resonance. 18

SQL Structured Query Language. 74

STD-NMR saturation-transfer difference NMR. 19

SVM support vector machine. 130

TSA thermal shift assay. 18, 19

Chapter 1

Introduction

1.1 Drug Discovery

The discovery of new drugs is a multivariate problem. A drug is required to bind to a target strongly and selectively, whilst having physicochemical properties that allow it to reach its target and remain at a high enough concentration for therapeutic effect. The strength of a compound's binding is referred to as its affinity, with factors important for maintaining therapeutic concentration referred to as the absorpotion, delivery, metabolism, excretion and toxicity (ADMET) properties. Changes to a compound's structure can lead to changes in its activity, known as the structure-activity relationship (SAR), however ADMET properties can also be affected by this change.

As both technologies and our understanding of disease improved, drug development evolved and changed the way that affinity and ADMET properties were considered. Before *in vitro* assays became common, use of *in vivo* testing meant that efficacy was being measured directly, implicitly accounting for target tractability and ADMET properties [3]. As focus moved to testing affinity *in vitro*, with ADMET considered separately, it became the responsibility of the medicinal chemist to balance these properties. The work flow shown in figure 1.1 gives an overview of the process of a clinical candidate's discovery.

1.1.1 Early Drug Discovery

As we entered the second half of the 20^{th} century, drug discovery was markedly different from today. Most notably, *in vivo* models were used for the primary screen [3], and projects relied on medicinal chemists synthesising compounds in gram quantities for use by pharmacologists.

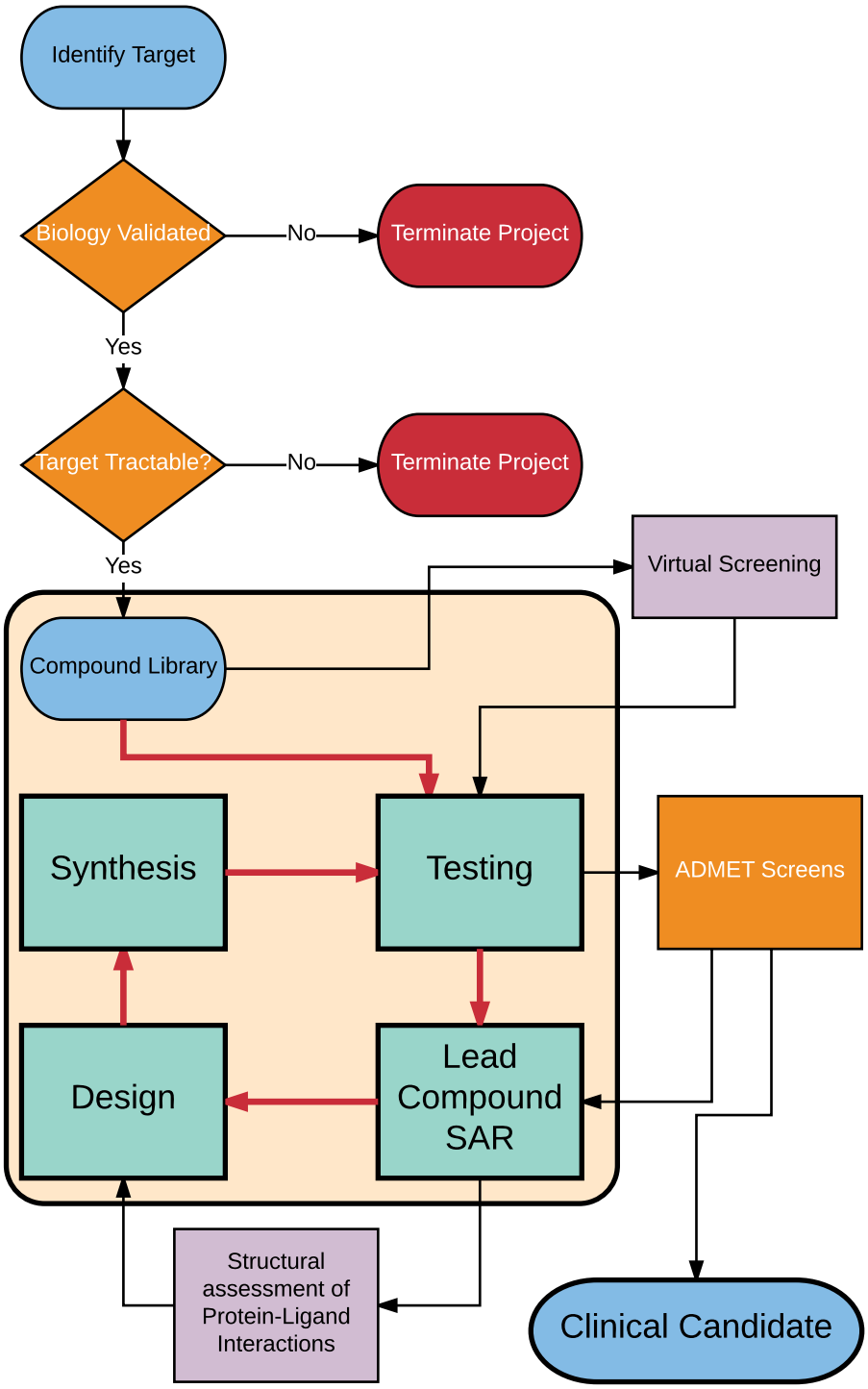


Fig. 1.1 Drug discovery work flow. The boxed region and red arrows show the steps taken in early drug discovery. Orange nodes represent processes that were considered implicitly in early drug discovery through the use of *in vivo* models, and lilac nodes represent optional steps. Based on a figure from [2]

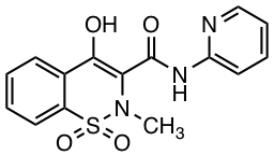


Fig. 1.2 Structure of Piroxicam

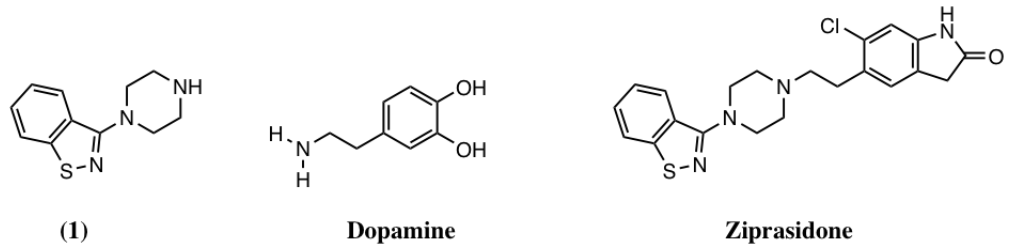
A project's starting point could be a serendipitous biological finding or an existing drug, as was the case for the discovery of piroxicam [4] as a treatment for arthritis. Known treatments included aspirin, ibuprofen and diclofenac - all carboxylic acids. They were rapidly metabolised and excreted, and as a result required multiple daily doses. This led to poor patient compliance and increased potential for toxicity.

Chemistry output was limited to a few compounds per week, therefore the design decisions made by medicinal chemists were vital. With the carboxylic acid group identified as the cause of metabolic liability, compounds with different acidic groups were synthesised. The development was guided by the compound's pKa and serum half-life in dogs. After several chemical families were tested over five years, oxicams were synthesised, ultimately leading to piroxicam (figure 1.2). Importantly, piroxicam was able to control the symptoms of arthritis from a single daily dose of 20 mg. Using *in vivo* models throughout development ensured that all ADMET properties were considered, however the total time from project start to approved drug was 18 years.

From the 1980s onwards, *in vitro* testing became more prevalent as the biological mechanisms of diseases were better understood. Ziprasidone, for the treatment of schizophrenia, is an early example of a successful drug that benefited from supplementing *in vivo* research with *in vitro* studies[5]. Having identified the dopamine type 2 receptor (D_2) as the target for known drugs, and that binding to the 5-hydroxytryptamine type 2 receptor (5-HT_2) is required to avoid unwanted side effects, Glennon et al. [6] searched for compounds with *in vitro* binding to 5-HT_2. Compound (1) in figure 1.3 was identified as a potent 5-HT_2 binder. By combining it with dopamine, the natural ligand of D_2, binding to both receptors was achieved. Further SAR and modification ultimately led to ziprasidone.

1.1.2 Modern Drug Discovery

Modern drug discovery has seen a move towards a reductionist approach, letting molecular and cell biology play a leading role. As we are no longer dealing directly with *in vivo*



(1) **Dopamine** **Ziprasidone**

Fig. 1.3 Ziprasidone

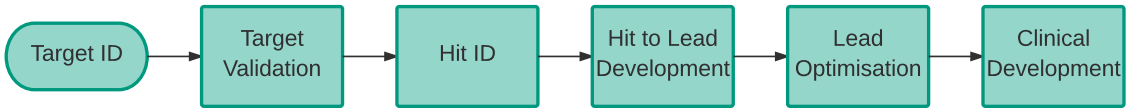


Fig. 1.4 A linear overview of the modern drug discovery work flow

efficacy, the work flow (figure 1.4) requires explicit identification and validation of target. Working with molecular targets rather than whole systems allows for much higher throughput experiments, leading to larger initial screens. If the structure of the target or closely related protein is available, these starting points can then be developed rationally using structure-guided approaches. Coupled with improvements in synthetic chemistry, the SAR landscape can be explored much more quickly, streamlining the discovery of high affinity ligands. However, strong binding affinity alone does not yield a drug, and physicochemical properties must also be considered.

1.1.3 Target Identification, Validation and Tractability Assessment

For a drug to be efficacious, binding to its target protein (or other macromolecule) must result in alteration of a biological process in such a way that modifies a disease[7]. A lack of efficacy is a major cause of failure, often arising during the very expensive clinical stages, after large investment in the development of the clinical candidate [8]. Better identification of disease modifying targets could therefore improve success rates and allow early termination of unsuccessful projects [9].

There is a range of techniques for target identification, which can be grouped into two distinct strategies: molecular and systems approaches[10]. The systems approach identifies targets through the study of disease in whole organisms, making use of data from *in vivo*

studies. Since the start of the 21^{st} century, the molecular approach has become the primary method for target identification. It aims to understand cellular mechanisms, and therefore makes use of clinical samples and cell models of cells implicated in the disease[10].

Modern research generates large amounts of data. In the context of target identification, bioinformatics can be used to prioritise disease targets [11]. Many databases have been created to provide information on known drugs and their targets [12–16], and most recently the Open Targets platform extends this to potential targets [17], making target identification data available for new projects.

Once a target is identified, it can be studied further using a range of validation techniques to suppress production of the protein as proxy for inhibition. One approach is to use antisense technology- RNA-like oligonucleotides designed to be complementary to a region of the target mRNA molecule[18]. Binding to the mRNA then prevents the synthesis of the target protein. Another approach is to use gene knockouts, where animals have had the target gene deleted or disrupted to prevent protein expression. Although powerful, animal models are sometimes difficult to develop for certain disease types, such as psychiatric illness and stroke [19].

Recent work by Hingorani and colleagues (preprint)[20] suggests that population genetic association studies can be used as a "natural randomised trial." Presence of germ line genetic variants, which cause a change in the expression or activity of a protein, is analogous to receiving drug treatment in a randomised controlled trial (RCT) performed in phase 3. Normally for novel targets, the phase 3 RCT is the first test within humans to see whether it has an effect on the disease, meaning targets are not truly validated until the final stages of the expensive drug discovery process. This approach would highlight efficacious targets at the start of the process and greatly reduce the number of failed projects, however, it will require datasets of genomes annotated with clinical data, raising legal, ethical and data protection issues. Although this approach could result in better early target validation, further questions remain regarding whether the target is suitable for treatment with a small molecule.

Drugs are usually administered in tablet form, meaning that they must be able to reach their target starting from the gastrointestinal tract. Oral bioavailability requires drugs to be soluble in aqueous solution, but also to be permeable through cell membranes. Lipinski *et al* [21] developed the rule of 5 to show the physicochemical properties required for good oral bioavailability.

Poor absorption or permeation are more likely when:

- There are more than 5 H-bond donors (expressed as the sum of OHs and NHs)
- There are more than 10 H-bond acceptors (expressed as the sum of Ns and Os)
- The molecular weight is over 500
- The Log P is over 5
- Compound classes that are substrates for biological transporters are exceptions to the rule

In 2002, Hopkins and Groom [22] recognised that some binding sites would be incompatible with the properties required of orally bioavailable small molecules. As a result, only a subset of the human genome is capable of binding drug-like molecules, leading to the concept of the "druggable genome". The intersection between the druggable genome (~3,000) with disease-modifying genes (~3,000) give potential drug targets, estimated to be between ~600-1500, as shown in figure 1.5. A more recent analysis by Finan and colleagues [23] has revised the estimated size of the druggable genome to be 4,479. This increase can be partly attributed to the inclusion of proteins targeted by biotherapeutics such as monoclonal antibodies, which can bind to proteins that are incompatible with small molecule binding.

Usage of the word druggability has varied in the literature since its introduction, also prompting the more recent term "ligandability" [24]. Hopkins and Groom defined a target as druggable if it was able to bind orally bioavailable druglike molecules, but stated that "Druggable does not equal drug target". Hajduk and colleagues' [25] use of druggability was simpler still, describing the target's ability to bind a small molecule with high affinity . Cheng *et al.*'s definition of druggable [26] required modulation of the target, introducing the need to also have a functional effect. In 2011, Edfeldt and colleagues [24] recognised that there were varying usages of the term "druggability", and introduced the term "ligandability" to take the definition of "able to bind a small molecule with high affinity." Bauer and Breeze [27] provide a particularly in depth description, wrapping up all target requirements of a successful drug discovery project in the single term.

"Tractability" will be used here as an umbrella term for all of these measures, table 1.1 summarises the different tractability terms and how they are used in the literature. The measures have been given the name used in the publication, but have also been assigned a category that matches the definition type. The three categories used are "Ligandability",

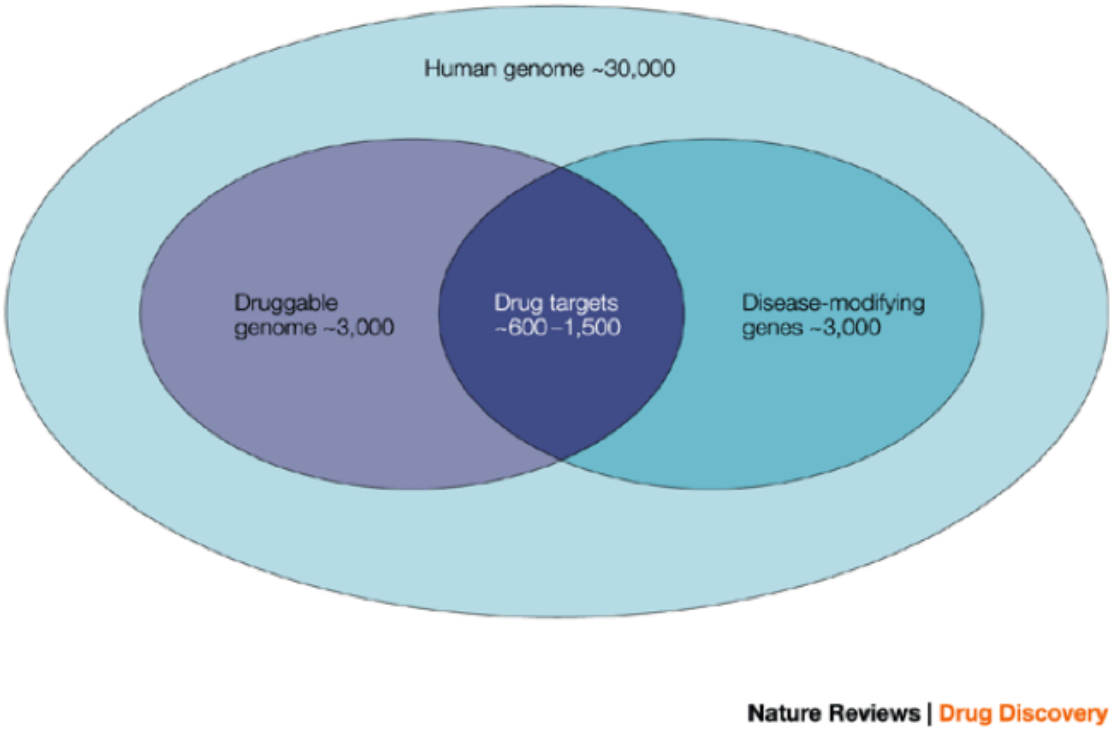


Fig. 1.5 Venn diagram showing the predicted number of drug targets, taken from Hopkins and Groom [22]

"Druglike ligandability" and "Drug Feasibility". Ligandability matches the Edfeldt definition, druglike ligandability has the added requirement of binding a ligand with druglike properties, and drug feasibility considers all factors a target needs to yield a drug.

Here, the drug feasibility definition of druggability will be used. Given this definition, it is difficult to determine the druggability of a target until the advanced stages of a project, or if a drug molecule is already known. This definition of druggability can be split into two components: ligandability and target quality (figure 1.6a). Target quality addresses the difficult questions such as a target's impact on disease, and most of this assessment is performed during the target identification and validation stage.

Figure 1.6b extends the two dimensional plot to also include ligand development. The length of each arrow roughly represents the amount of investment put into a project. The "Borderline Success" arrow represents a target that shows a moderate ligandability and target quality. Ligandability assessment is available early on in a project in the form of both experimental and computational approaches. In this example, the borderline case has met the ligandability criteria and will be progressed, however a target with comparable target quality and lower ligandability will be abandoned early.

Due to the complex nature of target quality, we can assume a large error along the x axis. If a target was predicted to be a borderline success and continued on in its development, it may become apparent late on in the project that it in fact had a much lower target quality [23], typically during the more expensive clinical trials stages.

Protein-protein interactions (PPIs) are considered the "high hanging fruit" in drug discovery [28], and projects are pursued even where ligandability is thought to be low, but the target quality is high. Many ligandability measures attempt to predict a pocket's ability to bind leadlike molecules from high throughput screening (HTS), however PPIs often perform badly in high throughput screens [29, 30]. PPIs interact across large and relatively featureless sites [31], lacking the topologies such as large pockets, clefts and groves normally targeted by small-molecule inhibitors. However, they do contain hotspots [32], regions of disproportionately high affinity that can act as footholds. Although PPIs are considered unligandable by many predictive methods, they are particularly suitable for fragment-based drug design (FBDD) [33, 24], which will be discussed further in section 1.2. Hotspots also exist within more traditional small-molecule binding sites, showing that the ligandabilty of a pocket does not rely on features of the pocket as a whole, but regions within the pocket. The method

Table 1.1 Tractability measures from the literature categorised by scope

Authors	Year	Measure	Category	Definition
Hopkins and Groom	2002	Druggability	Druglike ligandability	Target's ability to bind orally bioavailable molecules. Druggable does not equal drug target [22]
Hajduk *et al*	2005	Druggability	Ligandability	Target's ability to bind a high-affinity ($K_D < 300nM$, non-peptide, non-covalent inhibitor) [25]
Cheng *et al*	2007	Druggability	Drug Feasability	Likelihood of modulating a target by oral small-molecule drugs[26]
Edfeldt *et al*	2011	Ligandability	Ligandability	The ability of a protein target to bind small molecules with high affinity. Although ligandability is a requirement for finding drugs for a particular target, it is not a guarantee that such ligands will make good drugs
Bauer and Breeze	2016	Druggability	Drug Feasability	The amenability of a molecular target (within the context of a cell, tissue, or whole organism) to pharmacologically useful functional modulation by synthetic compound(s) with drug-like properties. Although conceptually attractive, druggability is in practice a composite parameter that encompasses not only target binding affinity but also cell permeability, serum stability, oral bioavailability, and other pharmacokinetic, pharmacodynamic, and toxicological properties of the molecules that are found to interact with the target [27]

described within this thesis goes beyond providing an assessment of the pocket as a whole, and is able to specifically identify the hotspots and the interactions that cause hotspots.

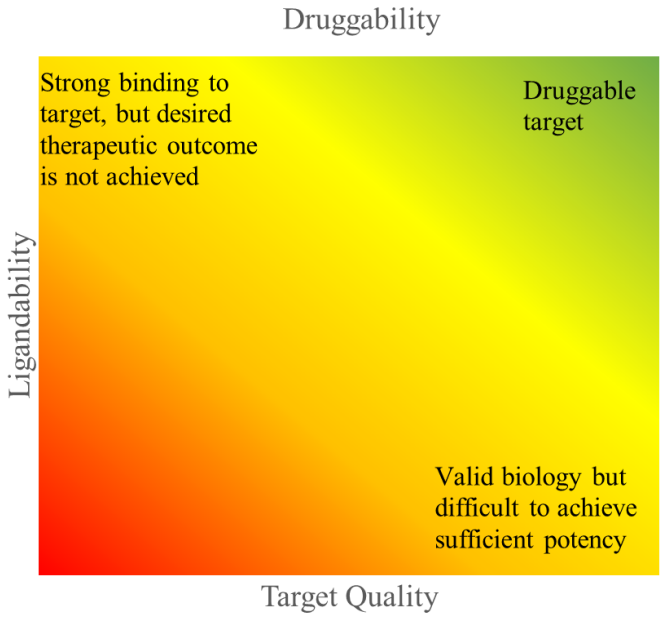
1.1.3.1 Current Tractability Assessment Methods

The most suitable method for assessing tractability for a given target will depend on the data available. Figure 1.7 summarises the different indicators of tractability, ranging from high confidence indicators such as the existence of an approved drug, through to the low confidence sequenced-based prediction models. However, the most attractive targets for drug discovery often have very little information available. Computational predictions provide the greatest utility in this case, and compared to the other indicators, such as "existence of high affinity drug-like molecule", have the greatest scope for improvement.

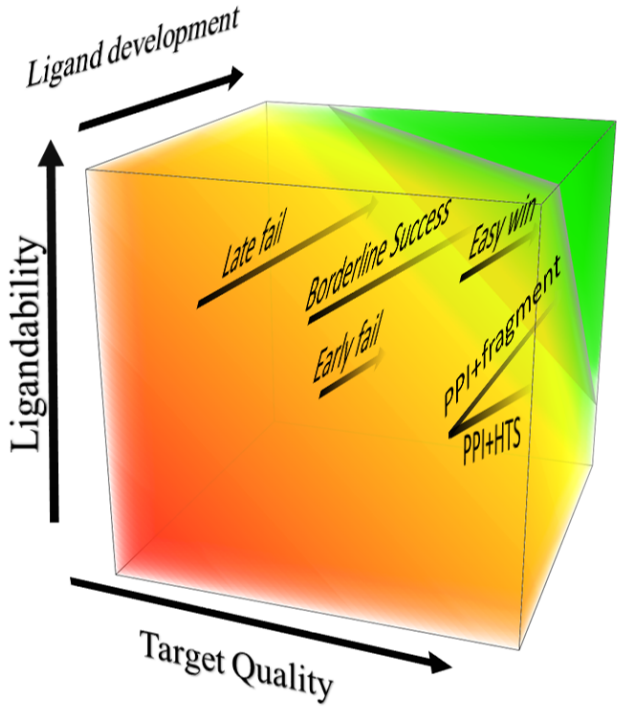
Hajduk *et al.*[25] were able to classify pockets as "druggable" or "undruggable" (ligandable or unligandable) by predicting the fragment hit rate using a regression analysis and 8 descriptors. The results of fragments screens for 23 proteins were examined, and a high correlation between the fragment hit rate and the likelihood of discovering a nanomolar inhibitor was found. Therefore a range of descriptors including apolar surface area, roughness and cavity volume were used to train a method that predicted the fragment hit rate. This was then able to classify 94% of known ligand binding sites as moderately or highly druggable.

Druggable pockets have properties that differ from undruggable ones. Schmidtke *et al.* published a freely available druggability dataset[34]. It used Cheng *et al.*'s definition of druggable (1.1), and a druggable cavity directory was created by cross referencing lists of oral drugs with the protein data bank (PDB) [35]. Analysis of this dataset showed druggable binding sites contain 20-40% polar surface compared to 40-60% for non-druggable sites. 70% of the polar atoms in druggable sites are found to have small solvent exposed areas, which decreases to 50% in undruggable sites. In druggable sites the polar atoms are surrounded by a hydrophobic region, and are found to protrude into the cavity, making them more available for interaction.

Datasets of druggable proteins have been used to create computational druggability prediction methods. Cheng's dataset was used for SiteMap[36], whereas a more recent program, DogSiteScorer[37], extended the Cheng dataset with Schmidtke's[34] and Hajduk's[25] datasets. In all of these cases, machine learning was used in conjunction with multiple descriptors to differentiate between druggable and undruggable pockets. A disadvantage of this approach is that the pocket is considered as a whole, and loses fine details such as



(a) Druggability split into its two components of target quality and ligandability. The green area represents a druggable target



(b) Including ligand development, an arbitrary measure of time and money, on the z axis. Each arrow represents a drug discovery project, with the start of arrow representing the target quality and ligandability (i.e. the druggability) and the head representing the outcome. The green region represents a successful project

Fig. 1.6 Relationship between target quality, ligandability and druggability

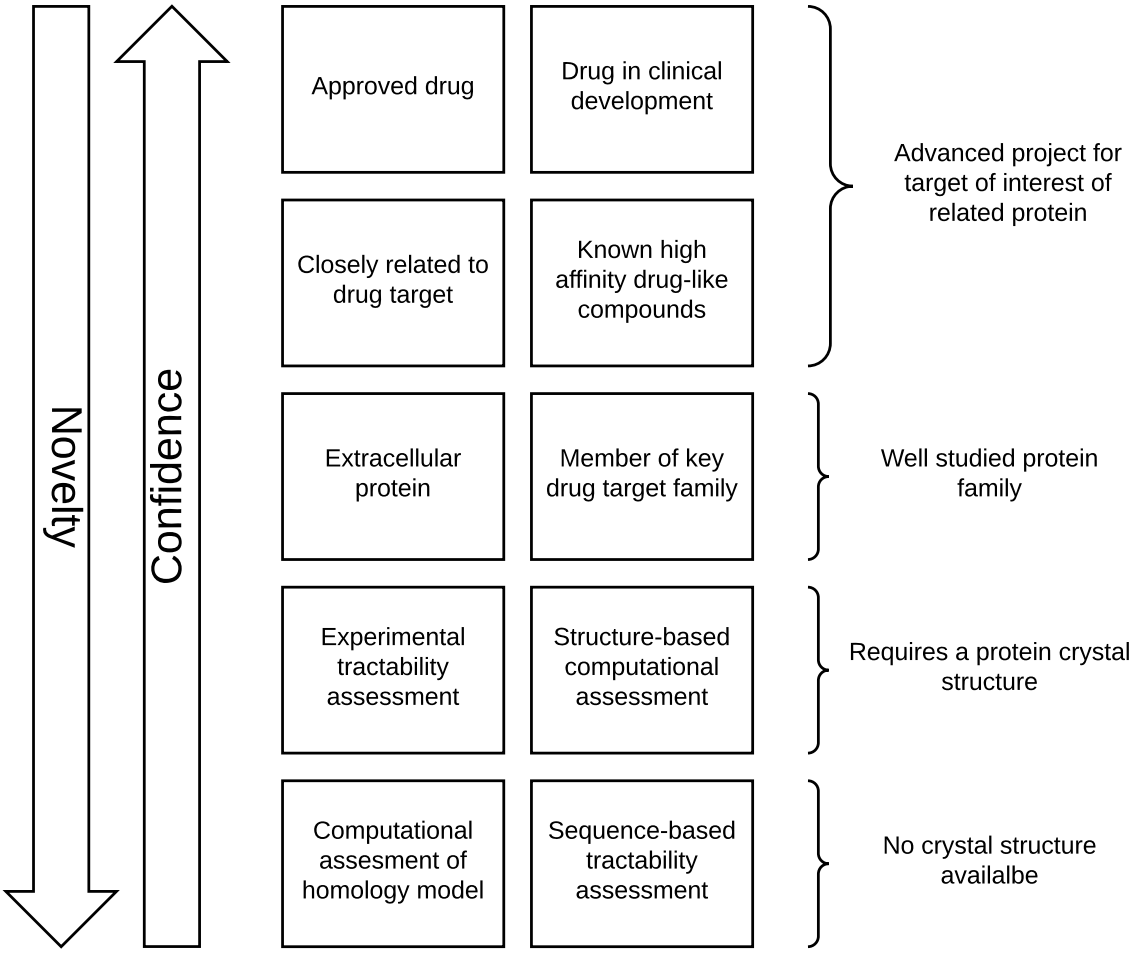


Fig. 1.7 Overview of protein tractability assessment

the location of the hotspot. This means that such a method is unsuitable for the finer detail prediction of fragment binding sites.

1.1.4 Hit Identification

1.1.4.1 Computational

In modern drug discovery, most projects are backed up by computational support. computer-aided drug design (CADD) has two main branches: molecular modelling and chemoinformatics, both of which are able to able to assist in hit identification.

Molecular modelling is a broad term covering techniques that predict molecular behaviour, structure and properties. In the context of hit identification this means using known proteins, protein-ligand complexes and/or small molecules to suggest which compounds are likely to bind to the target, a process referred to as virtual screening. Virtual screening techniques filter or rank large virtual libraries to prioritise which compounds should be tested experimentally. Although virtual screening faces challenges, explored in detail in chapter 5, it can to provide an enrichment of active molecules in a subset of a larger library.

Chemoinformatics is the use of informatics to solve chemical problems[38]. The methods used are typically fast, and applicable to databases of molecules. Molecules can be represented in numerous ways. Fingerprints, or binary string representations, are a highly efficient description molecule. A series of 1s and 0s represent the presence or absence of a certain chemical feature, and these fingerprints can be rapidly compared to show their similarity. The application of this method to virtual screening will be discussed further in chapter 5. Simplified molecular line input entry system (SMILES) notation[39] and InChI[40] provide two linear notations that can be used to generate the 2D structure of compounds. The CANGEN algorithm[41] can be used to create canonical SMILES representations for compounds, useful for assessing whether there is an exact match between two molecules.

In addition to finding similar molecules to known actives, chemoinformatics can aid hit identification by identifying a representative subset of library. Clustering groups together similar compounds, and picks one as a representative of that cluster. If a chemical database contains 1,000,000 molecules, but screening capacity is limited to 100,000, clustering can provide the most diverse subset of the database, providing the greatest coverage of chemical space.

1.1.4.2 High Throughput Screening

During the 1990s, combinatorial chemistry and parallel synthesis led to large libraries of compounds [42–44], and automation of assays with fluid handling robots[45] provided a means to test them - HTS. The Human Genome Project's mission to squence the human genome was underway[46], and by the mid 90s was able to identify genes involved in disease. With so many targets and compounds to screen, the screening technology was seen as partially limiting[47]. HTS attracted a lot of investment, as pharmaceutical companies set to capitalise on this opportunity. Organisations using HTS previously able to screen 20 targets per year with a library of around 75,000 compounds could screen 100 targets with one million compounds by the late 90s.

While previous methods aimed to make use of existing knowledge to discover hits, HTS takes an intellectually neutral approach. Emphasising large libraries and fast experiments, HTS relies on active molecules being present amongst the hundreds of thousands in the screen. Initially it was hoped that it would be possible to derive drugs directly from an HTS screen[48], however even with multi million compound screening decks, the output is limited to chemical starting points in need of optimisation. This limitation can be explained by the size of leadlike [49] chemical space, which represents all potential molecules with leadlike properties. It is estimated that there are more than 1×10^{30} possible molecules[50] and even the largest HTS screening deck represents a tiny fraction of this space.

Early libraries were put together based on the previous activities of the company, with little consideration of the suitability of the compounds. Furthermore, early combinatorial chemistry relied on a small number of reactions, leading to limited diversity. As the success of HTS relies on how well leadlike chemical space is covered, companies made large investments to improve this coverage. First, collections were cleaned up by removing compounds with undesirable properties, with some libraries having as much as 40-50% of compounds removed[51]. Following this, there were large investments to increase library size with carefully selected compounds. Chemoinformatics was used to ensure diverse chemotypes were selected, whilst also considering leadlike properties. Although these later HTS libraries still showed poor coverage of lead-like chemical space, this optimisation resulted in a better representation of this space. A more recent approach with a far greater coverage of chemical space, called FBDD, will be discussed later in section 1.2.

1.1.4.3 Focused Screening

An alternative to trying to maximise coverage of chemical space is to focus on a relevant region of chemical space. Libraries can be designed to interact with either an individual target or a family of related targets. Libraries can be created by using known actives of similar binding sites[52], or rationally designed using *in silico* methods[53]. Simply recycling known actives is likely to run into intellectual property (IP) problems, therefore it is important to combine information about known ligands with structure-based methods to modify side groups around key cores. This approach is less likely to identify novel starting points for a drug discovery project, but offers a more efficient approach compared to HTS.

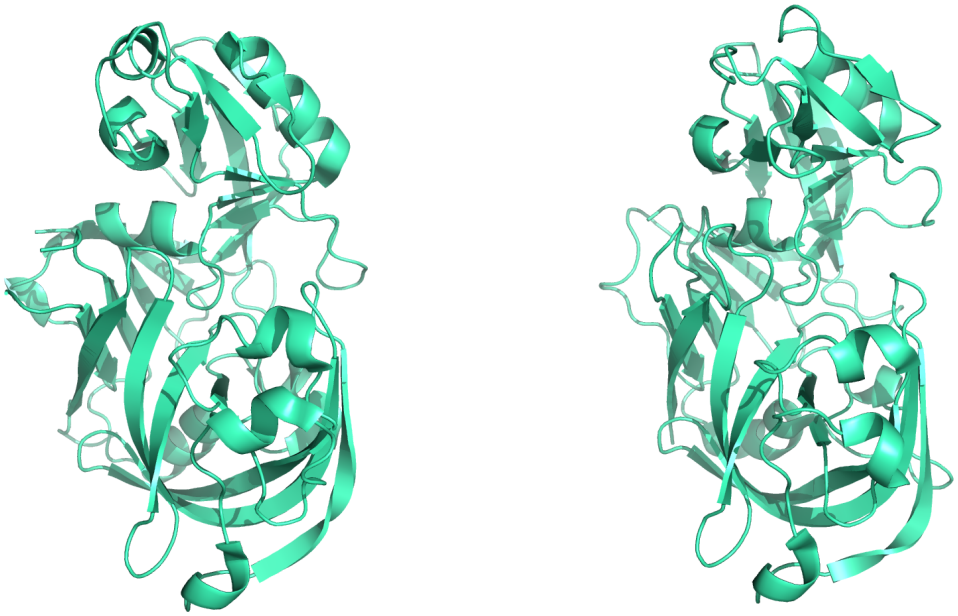
1.1.5 Hit-to-Lead Development

1.1.5.1 Structure-Guided

By the mid 1980s, drug discovery projects began to make use of protein structures to rationally design molecules. In cases where the target's structure was unavailable, models could be created based on homologues for which a structure was available. An example of this was the use of endothiapepsin's structure (figure 1.8b) with the program FRODO [54] to create a model for renin[55] (figure 1.8a), a target of interest in the search for a new treatment for hypertension. The similar 3D shape of these models could be used by medicinal chemists to grow their molecules whilst maintaining shape complementarity between the protein and the ligand.

The manual process of creating a model from a related structure was initially viewed with scepticism from the protein crystallography community [56], therefore Blundell and colleagues set out to automate this process to make it more widely accessible. The first program, COMPOSER [57, 58], assembled fragments of structures from homologous proteins. A later program, MODELLER [59], used spatial restraints based on knowledge from a related protein, and became widely used in both academia and industry.

In time, it became more common to have a protein structure available for your target of interest. The scope of a structure's utility in a project extended beyond simply guiding the growth of known binders, with application much earlier on. Protein structures were used to assess the tractability of a target, providing information such as whether the target has a suitably sized pocket available for binding. In addition to visual inspection of protein



(a) Renin **(b)** Endothiapepsin

Fig. 1.8 Cartoon representation of two aspartyl proteases

structures, computational methods were being developed to interrogate this information rich resource.

GRID was developed in 1985 by Peter Goodford to computationally map protein binding sites with favourable regions for a given probe. As the name suggests, this method produces a grid output, which can be used visualised as 3D contours over the protein structure, guiding medicinal chemists.

Improvements in computational power and methods allowed structure-based approaches to be used in hit identification. The most common structure-based computational approach is molecular docking. DOCK was developed in 1982 [60], initially used for binding mode prediction, vitally important for structure-based drug design. Ten years later, Kuntz described how DOCK can be used to computationally assess hundreds of thousands of compounds, now referred to as virtual screening.

1.1.5.2 Trends

One of the key aims during hit-to-lead development is to improve affinity for the target, however this needs to be balanced against physicochemical properties that may lead to ADMET issues. In the late 2000s, it was widely recognised that a decrease in drug discovery

productivity was partly due to poor quality compounds entering clinical trials [61, 62], with Leeson *et al* blaming the increased lipophilicity of compounds leading to promiscuity [61]. In 2011 Walters and co-workers [63] showed that between 1959 and 2009, molecular properties of lead compounds were diverging from those of known drugs, becoming larger, flatter and more lipophilic.

Keseru and Makara looked at the properties of hits and their subsequent leads from various screening methods [62]. They found that hits from HTS had worse properties than those discovered from fragment-based screening (discussed below) or natural products. Despite this, the properties of subsequent leads were the same, regardless of the screening method, demonstrating a tendency to gain potency through hydrophobic interactions during hit-to-lead. This demonstrates a need for careful optimisation of multiple parameters rather than affinity alone.

1.2 Fragment-Based Drug Discovery

Fragments are molecules that typically have a molecular weight between 120 and 250 and 8-18 heavy atoms. They follow a rule of 3 [64], analogous to Lipinski's rule of 5 [21]. Compared to high throughput screening, fragment-based drug design has an emphasis on efficiency, screening fewer compounds and synthesising fewer compounds in hit-to-lead. FBDD tends to ultimately result in compounds with more desirable properties[65].

Fragments are particularly good at exploring binding sites due to their low complexity[66]. As the size (and therefore complexity) of the molecule increases, the likelihood that it will complement the binding site decreases, however when complementarity is achieved the binding affinity is higher. Therefore more simple molecules will complement the binding site more often, but at an affinity often too low to be detected by traditional assay techniques. This suggests an optimal size for maximising useful binding events, where you have complementarity between the protein and the ligand at a detectable affinity. This demonstrates the importance of designing a suitable fragment library to match the sensitivity of your chosen screening technique.

$\Delta G_{Binding}$ has been shown to be related to the number of heavy atoms[67], and as fragments typically have fewer than 15 heavy atoms, they have a limited potential for binding compared to larger drug like molecules. Hotspots are sites able to interact efficiently enough with fragment molecules to overcome the limited number of interactions[68], and once the

fragments are developed further into drug-like molecules, the moiety corresponding to the original fragment is very sensitive to modification.

Ligand efficiency[22] provides a means to compare the affinity of different sized molecules. Aiming to improve ligand efficiency rather than affinity discourages gaining potency relatively easily by adding large lipophilic groups. Ligand efficiency is defined in equation 1.1.

$$LE = \frac{\Delta G}{Numberof HeavyAtoms} \tag{1.1}$$

This idea was developed further by the introduction of Group Efficiency[69], shown in equation 1.2. Group efficiency helps medicinal chemists to decide whether the addition of a group gives a sufficiently large increase in affinity, providing a guideline to suggest how much potency should be gained as a function of the number of heavy atoms added.

$$GE = \frac{\Delta\Delta G}{\Delta Numberof HeavyAtoms} \tag{1.2}$$

1.2.1 Screening Methods

Fragment screening typically starts with a cascade of assays[70–72, 68], starting with high throughput methods such thermal shift assay (TSA) [73–75] or surface plasmon resonance (SPR) [76, 77], with hits followed up by isothermal titration calorimetry (ITC), nuclear magnetic resonance (NMR) spectroscopy[78] and X-ray crystallography[79, 80]. ITC gives high quality information about the free energy, entropy and enthalpy of binding. NMR can show which residues interact with the ligand, and the X-ray crystal structure that can show the exact binding mode of the fragment.

It is also possible to bypass filtering with thermal shift or SPR, and use X-ray crystallography as an initial screen [81, 79]. Nienaber and colleagues [82] created cocktails of up to 100 fragments, which can be differentiated by their electron density, to allow screening of thousands of molecules per day. A disadvantage of such a large number of fragments within a cocktail is the low concentration of each individual fragment, important for the detection of weakly bound, but useful hits. Improvements in library design, cocktail design and soaking methodologies led to changes in screening methods [83]. Astex pharmaceuticals employs this approach [84], using cocktails of five diverse fragments that can be not only

easily differentiated by their electron density, but also reduce the likelihood of multiple fragments binding.

Schiebel and colleagues [85] compared six different screening methods to X-ray crystallography, looking at endothiapepsin. Previous attempts [86, 87] to screen with a fluorescence-based high concentration biochemical screen (HCS), saturation-transfer difference NMR (STD-NMR), reporter-displacement assay (RDA), native mass spectrometry (MS), microscale thermophoresis (MST) and TSA had shown very little overlap. While two thirds of the library were detected as potentially binding, only 41 out of 361 fragments were identified by two or more methods, and no hit was found in all six assays. Due to poor overlap, crystallographic screens were performed using single compound soaks of all 361 fragments. The authors were able to find 71 hits (20% hit rate), 31 of which had not been found by any of the biophysical assays, and a further 21 were only predicted by one assay. In this example, the use of a cascade with two or more screening techniques would have failed to identify 73% of the crystallographic hits.

The authors note that the fragments found by biophysical assays only bound close to the catalytic dyad, whereas fragments identified by crystallography alone found 11 binding sites in total [88]. They continue to say that the binding sites from fragments identified in the biophysical assays present the best starting point for fragment growth, whereas the additional sites found in the crystallographic screen provide important structural information that can be used for fragment growing. An overlay of all 71 fragments is shown in figure 1.9. Neighbouring proteins in the crystallographic environment have been included to highlight the fact that some of the extra fragment-binding sites can be an artefact of crystal contacts. The hotspot predicted by the method described within this thesis, Fragment Hotspot Maps [1], has also been displayed. The predicted hotspot coincides with the fragments highlighted by the authors as the best starting points for fragment growth.

Recent work by Frank von Delft and colleagues represents the cutting edge of high throughput crystallography. Their highly automated XChem facility can screen up to 1000 compounds individually within a week [89]. A recently developed method, PanDDA [90], utilises the large amount of data produced by XChem to provide a much clearer view of ligand electron density. Their method uses the electron density from the many *apo* structures, which are generated in cases where fragments do not bind, as a way to remove noise. As a result, they are able to locate many more fragment binding sites, even when they are only partially occupied. While this work represents an important step forward in protein crystallography,

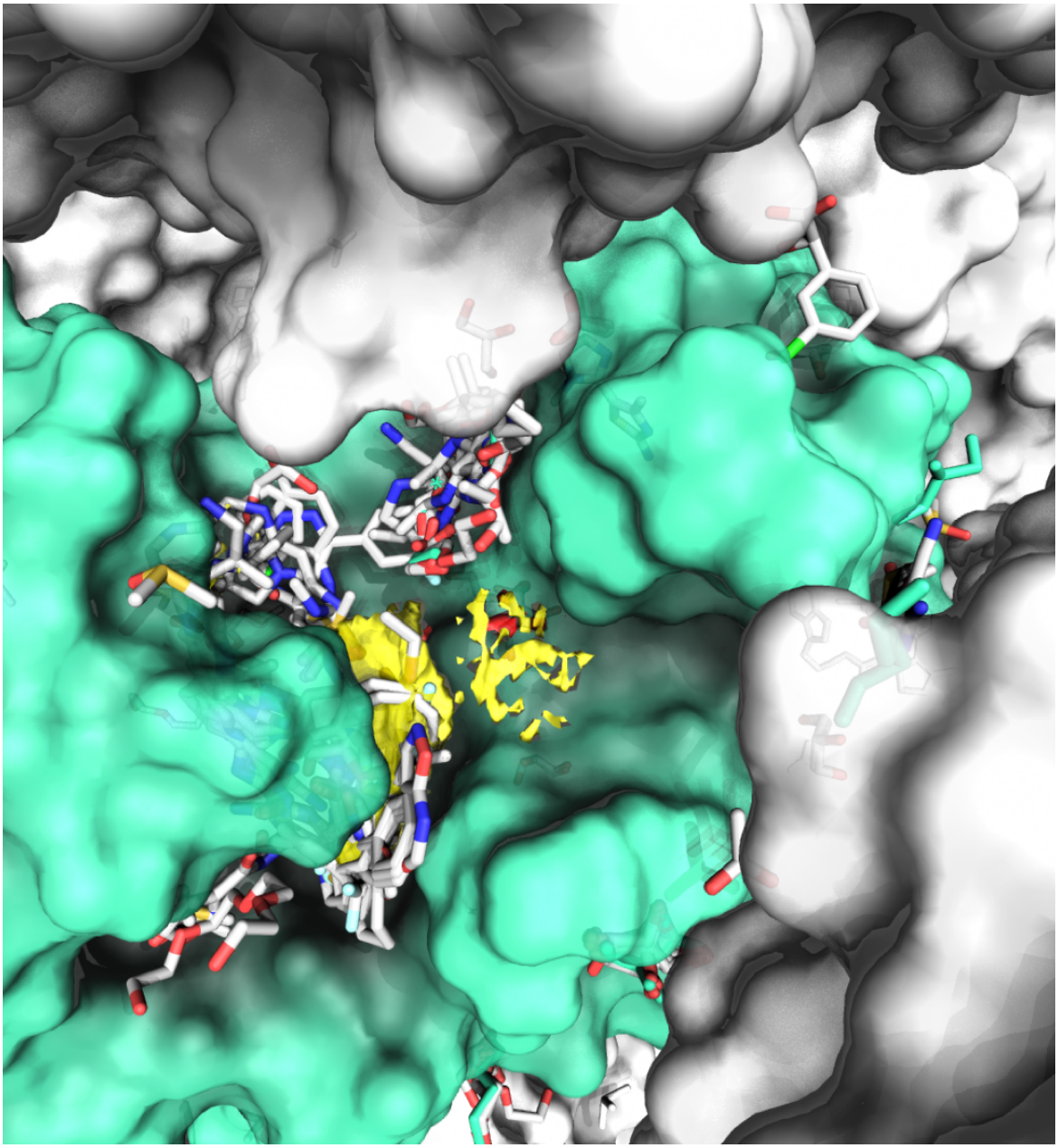


Fig. 1.9 Endothiapepsin displayed with 71 fragments. Fragment Hotspot Maps [1], introduced in chapter 2, are displayed as yellow (apolar), red (hydrogen bond acceptor) and blue (hydrogen bond donor) surfaces. Ligands are displayed as white sticks, endothiapepsin is displayed as a green surface and neighbouring proteins in the crystallographic environment are displayed as a white surface

it also presents a challenge. Much like the endothiapepsin example above, not all of these fragments will be suitable starting points for development, but offer important insight that can guide the hit-to-lead development. The Fragment Hotspot Map method, described within this thesis, may provide a computational approach to help categories fragments as those that will make good starting points, and those that can be used as information to guide the subsequent fragment development.

1.2.2 Fragment Development Strategies

Once a fragment hit is identified, it needs to be developed further to gain sufficient affinity. Despite starting from millimolar potencies, it is possible to obtain a nanomolar lead compound through the synthesis of 20-100 molecules from the the starting fragment hit [91]. This is made easier if knowledge of key interactions is available from X-crystallography or NMR, although a recent poll on the practical fragments blog http://practicalfragments.blogspot.co.uk showed that just under half of respondents (143 total) would begin fragment development without structural information (figure 1.10). Although it is possible to develop fragments without a structure [92], I will focus on structure-guided methods as they offer the best opportunity to develop efficient leads, and are the most relevant to this thesis.

There are three main strategies available for the development for fragments; linking, merging and growing. Fragment linking was introduced as the first fragment based approach in SAR by NMR by Shuker and colleagues [78]. Two fragments with micromolar affinities were tethered together to create a nanomolar inhibitor. Fragment linking was supported by computational approaches such as CAVEAT [93], HOOK [94] and CONCERTS [95], which aimed to link two proximal molecules in the binding sites. A requirement for successful fragment linking is the design of an optimal linker that is capable of maintaining the binding positions of the two fragments. Despite the support of computational approaches, this has proven difficult in practise, and fragments are susceptible to movement during the linking process [96].

Fragment merging potentially faces the same pitfalls as linking. However, if two fragments show suitable overlap with a common group or ring, they may be combined to form a single molecule. This has been demonstrated by Nikiforov and co-workers [97] in the search of inhibitors of the *Mycobacterium Tuberculosis* target EthR. Two micromolar fragments showed partial overlap, and when combined led to improved affinity and retained the binding position of the two parent fragments.

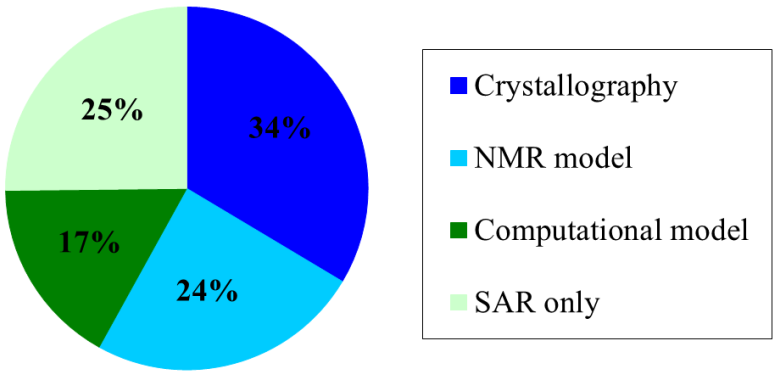


Fig. 1.10 Break down of 143 responses to the question: "how much structural information do you need to begin optimising a fragment?" From the practical fragments blog written by Dan Earlanson http://practicalfragments.blogspot.co.uk/2017/06/poll-results-what-structural.html

Fragment growing is a more straightforward approach, as it only requires a single starting point. The orientation of the fragment in the binding site is used to grow the fragment and pick out additional interactions. This can be aided by docking, or simply through knowledge of the binding site [98]. Once again, it is important that the original binding position is maintained in the fragment upon elaboration. One might expect that as other interactions are introduced, the original fragment portion of the lead compound could be pulled away from its original position. This is not the case, and it is typical for the initial fragment to show very little movement during elaboration [99]. Fragments are suitable for elaboration if they are anchored in place at a binding hotspot. As this is not necessarily clear from the structure alone, computational methods can help to prioritise which fragment to take forward.

1.3 Hotspots

Many proteins have pockets that have evolved to bind small molecules, and within these pockets are hotspots; areas that make a disproportionately large contribution to binding affinity [32]. The idea of small molecule hotspots evolved hand in hand with fragment-based drug design (FBDD). In 1996, Karen Allen [100] and colleagues described their multiple

solvent crystal structure (MSCS) method, solving the structure of porcine pancreatic elastase with acetonitrile as the probe organic solvent. The authors noted that their work had a "theoretical counterpart," the multiple copy simultaneous search (MCSS) method described by Miranker and Karplus in 1991 [101]. However, while the computational MCSS found a large number of favourable binding positions, the experimental MSCS found that the probe molecules only bound to relatively few sites.

In the same year, Rejto *et al* [102] found that the pipecolinyl moiety of FK506 acted as a "molecular anchor" for binding to FK506 binding protein (FKBP-12). Recognising the importance of the location at which the pipecolinyl moiety bound, they addressed this by splitting FK506 into fragments. Each of these was docked into the binding site, and all seven of them bound to the hotspot, rather than rediscovering their original binding poses. Months later, Shuker and colleagues [78] published their work on "SAR by NMR". Compounds with nanomolar affinities were discovered by linking together two smaller fragments that were found to bind to proximal sites.

In 1999, the MSCS method was studied further in two separate papers [103, 104], both of which finding once again that there were far fewer binding sites than computational methods suggest, which do not take into account solvation and entropic effects. Two years later, another MSCS paper from English and co-workers [105] compared their experimental results with GRID [106] and MCSS. Once again, they found that only a handful of the predicted sites corresponded to experimental sites, this time making the connection to previous work by Clackson and Wells [107] describing hotspots at protein-protein interfaces. In their discussion, the authors make an important statement:

> ...a disparity in the predictions might be anticipated since entropic and solvation effects were not explicitly included in the calculations. In general, electrostatic interactions dominate the computational predictions as they tend to be overestimated *in vacuo*. This comparison serves to highlight the amphipathic nature of these probe molecules (particularly isopropanol and acetone), with the observed binding mode representing a compromise between hydrophobic and hydrogen bonding interactions.

This observation turned out to be a fundamental one. The computational method described within this thesis was designed with this observation in mind, and other successful hotspot detection methods can be attributed to the direct or indirect handling of this idea, which will be discussed further in chapter 2.

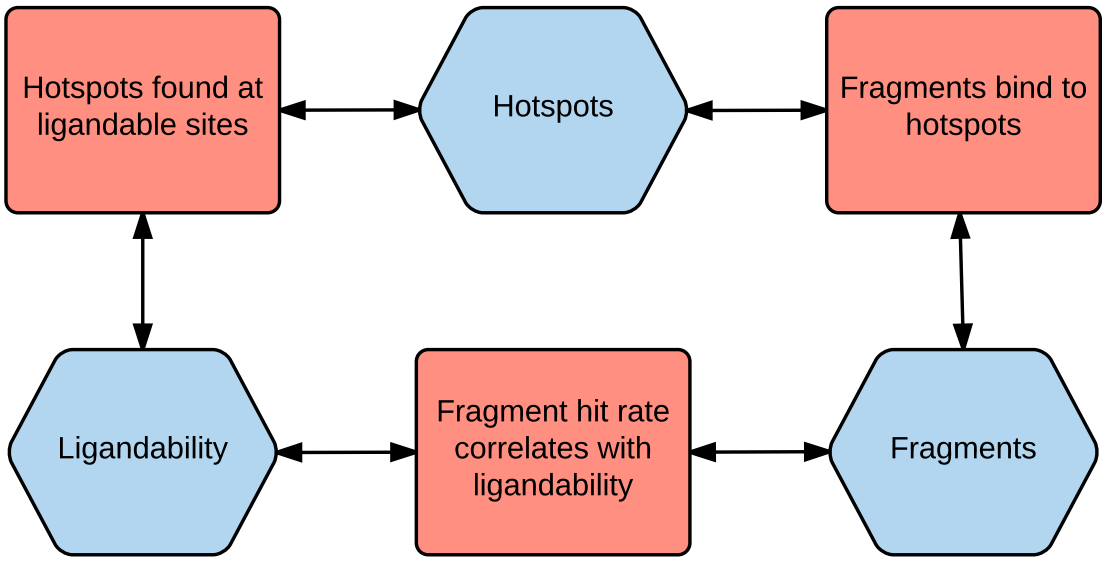


Fig. 1.11 Relationship between hotspots, fragments and ligandability. The red boxes show established ideas from the literature.

A second observation was that the identified probes could be used to generate pharmacophores for database searching, or to design more focused libraries for screening a given target. The latter idea provided a crystallographic alternative to the SAR by NMR approach, [78, 108], a precursor to crystallographic fragment screening [84].

Fragments, hotspots and ligandability are fundamentally linked, and it is this relationship that makes hotspot detection methods useful (figure 1.11). In 2005, Hajduk and colleagues [25] demonstrated a strong correlation between fragment hit rate and the ability to develop a high affinity inhibitor. Years later a similar analysis was performed within AstraZeneca [24], which took into account fragment hit rate, affinity and diversity of hits to score 36 projects from 2001-2008. All projects classified as "low ligandability" failed to yield HTS hits that entered hit-to-lead, although two of these found success with fragment-based approaches.

Multiple fragments typically bind to the same region of the binding site - the hotspot. In order for detectable fragment binding to occur, it must overcome two things: a limited binding interface [67], and loss of rigid body entropy [109]. Although the loss of rigid body entropy affects all ligands, the size of the penalty is mostly independent of size, and estimated at 15-20 kJmol^{-1}. Equation 1.3 shows the free energy of binding broken down into ΔG_{rigid}, the free energy associated with the loss of rigid body entropy upon binding,

and $\Delta G_{intrinsic}$, the remaining free energy terms that contribute to binding. Hotspots are capable of contributing enough to $\Delta G_{intrinsic}$ in a concentrated area to bind fragments [68]. Computational hotspot prediction methods have been used to predict both fragment binding sites and ligandability [110].

$$\Delta G_{total} = \Delta G_{intrinsic} + \Delta G_{rigid} \tag{1.3}$$

1.4 The Cambridge Structural Database

1.4.1 What is the CSD?

The Cambridge Structural Database (CSD) was founded in 1965 by Olga Kennard as a worldwide repository of small molecule crystal structures. The database is maintained by the Cambridge Crystallographic Data Centre (CCDC), who are responsible for curating new structures and providing access to the data. The CSD provides two key benefits to the scientific community. Firstly, it provides a single collection of standardised structures, allowing for easy sharing of high quality data. Secondly, the CSD can be studied as a whole. In 1997, Olga Kennard recalled her thoughts at the inception of the CSD [111]:

> The database was established in 1965 to fulfil a dream of myself and a great scientist, the polymath J.D. Bernal. We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments.

Both the number of structures within the CSD and the rate at which they are added are increasing each year, as shown in figure 1.12. Small molecule crystallography can achieve higher resolution than in proteins, and hydrogens are now routinely visible in the electron density. With the correct tools to access the data, the structural data in the CSD has become an invaluable resource.

1.4.2 Using Small Molecule Structural Data in a Drug Discovery Context

The vast amount of data in the CSD allows researchers to answer fundamental questions about molecular geometry and interactions. Access to molecular geometry data in the CSD

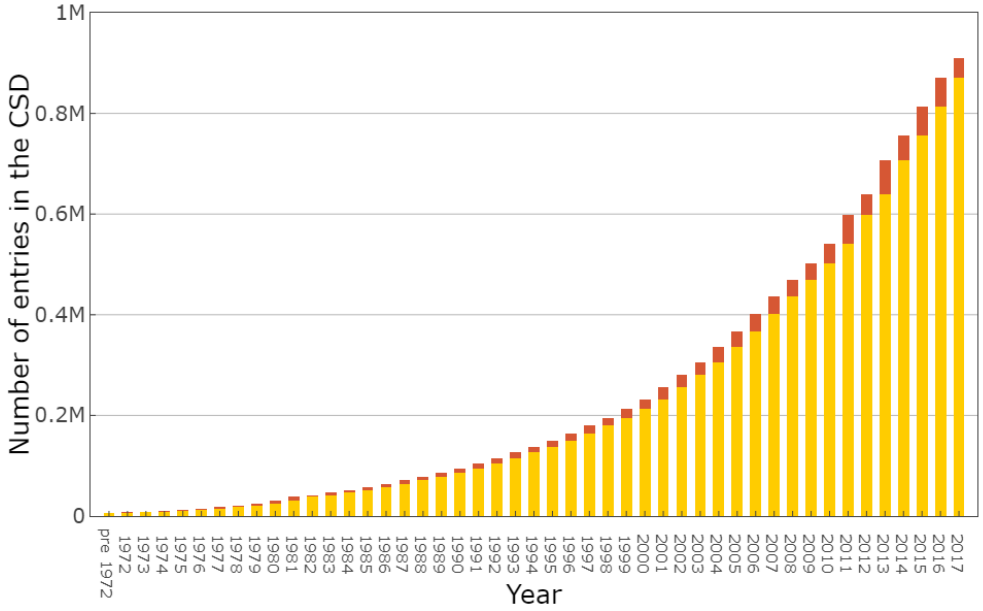


Fig. 1.12 Number of entries in the CSD by year. The red portion of the bar represents the number of new entries added that year

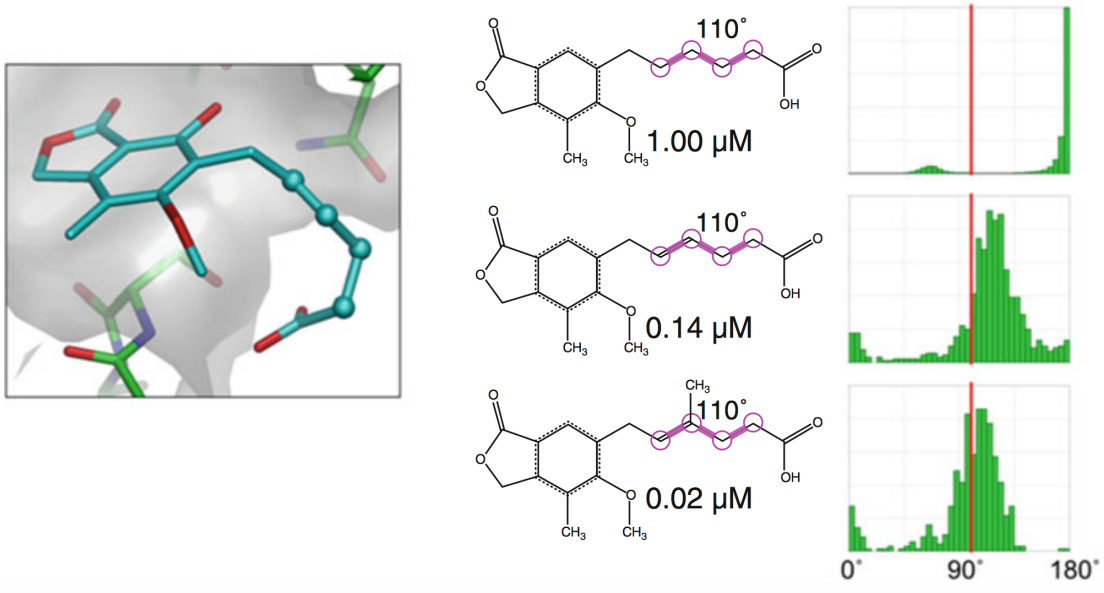


Fig. 1.13 Using mogul to aid the design of inosine monophosphate dehydrogenase (IMPDH) inhibitors. (Left) Strained IMPDH inhibitor. (Middle) Three structures with the key torsion angle highlighted. All three compounds have the same torsion angle once bound to the protein, but yield different affinities. (Right) Torsion angle distributions from the CSD. The vertical line represents the torsion angle required for for binding, and the histograms show the distribution of torsion angles for the given environment. Figure taken from original publication [113].

was greatly improved in 2004 with the introduction of Mogul [112]. This provided a simple interface for querying the CSD for bond-length, valence-angle, and torsion-angle distributions and statistics. This information can be used to optimise the conformational preference of a ligand such that it experiences minimal strain upon binding. This is exemplified [113] in figure 1.13, where three inosine monophosphate dehydrogenase (IMPDH) inhibitors of varying activity are shown. The slight changes to the compounds do not change the number or quality of interactions, and all three adopt the same pose, with a key torsion angle remaining at 110°. The difference is highlighted by the mogul distributions on the right hand side of figure 1.13, the required torsion angle is shown as a vertical line and in the original compound this is rarely observed in the CSD. Introduction of an allylic bond changes the preferred torsion from 180° to 120°, improving IC_{50} from 1 μM to 0.14 μM. Finally adding a methyl to one end of this bond moves the preferred torsion even closer to the 110°, and IC_{50} is improved further from 0.14 μM to 0.02 μM.

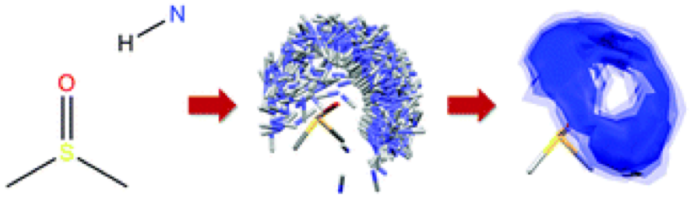


Fig. 1.14 Creation of SuperStar maps from IsoStar data. (Left) 2D diagram of dimethyl sulfoxide (DMSO) the central group, and NH nitrogen donor probe. (Middle) scatter plot of NH probes around the aligned DMSO central group. (Right) Blue surface showing isocontour of donor propensity, adapted from Wood *et al.* [116].

In addition to molecular geometry, and central to the work described in this thesis is the description of molecular interactions. Interactions in the CSD encompass all of the short range interactions between ligand and protein functional groups. There are two key pieces of software to explore interaction data in the CSD. The first is IsoStar [114], a library of molecular interactions that can be used to display the 3D distribution between a central group and contact group. All of the interactions between the central and contact groups are overlaid onto the central group, creating a "scatterplot" of contact group positions. This information is useful for assessing the interactions around a single functional group, however multiple functional groups can affect the same region of space, particularly in concave protein binding sites.

SuperStar was created in 1999 [115] to provide a clear way to map whole protein binding sites. It uses the data from IsoStar to create maps of propensity for a probe on a grid, a knowledge-based equivalent of GRID [106]. SuperStar breaks the input protein into fragments that correspond to IsoStar central groups. The chosen SuperStar probe, which will correspond to an IsoStar contact group, will be used to create an IsoStar scatterplot around the central group. These scatterplots are aligned to their original position in the protein, and each individual scatterplot is converted to a density map. These density maps are normalised and combined to provide a single description of the entire protein binding site. An overview of the process is shown in figure 1.14

A key feature of using a knowledge-based approach to describe interactions is that interaction types do not need to be defined, only probe types. If an interaction between a probe and central group is favourable, or more importantly competitive in comparison to other interactions [117], it will result in a signal. This is valuable as obscure but important interactions have the potential to be overlooked [118].

1.5 Aims

In this thesis I will describe a computational method for the identification of hotspots starting from just a protein structure. The Fragment Hotspot Maps method will provide a means of identifying fragment binding sites, providing information not only about where fragments will bind, but also which interactions are the most important. In order to best represent prospective use, the method will be run globally on apo protein structures. Once the method's ability to highlight fragment binding sites has been demonstrated, attention will turn to its utility in structure-based drug design.

Chapter two will first explore the literature that highlights the physical nature of hotspots, and gives a definition of a hotspot within the context of this project. The steps taken during a Fragment Hotspot Maps calculation will be described and justified in terms of these conclusions.

Chapter three will describe the method's validation, which considers hotspots as defined in chapter two. In addition to assessing the method's performance across the whole dataset, two examples with published group efficiency (GE) analysis are explored in detail. Finally, this chapter will look at how Fragment Hotspot Maps can be used directly to aid structure-based drug design, without any further analysis.

Chapter four will describe work that aims to improve the access to Fragment Hotspot Map calculations. An intuitive web server has been set up for people unaccustomed to running command-line tools, providing a means both to set up protein for the calculation and to visualise the results. In order to facilitate using the results for those more familiar with scientific programming, a Python-based Hotspot API has also been developed. This provides the ability to automate calculations over large numbers of structures, with functions to help use the results in wider work-flows.

Chapter five will demonstrate the ability of the Fragment Hotspot Maps to guide structure-based virtual screening. This used the Hotspot API described in the previous chapter, and two virtual screening methods, each run with and without the information from the Fragment Hotspot Maps.

Finally, chapter six will take a look at how Fragment Hotspot Maps are being used currently by collaborators. It will outline future plans and recent enhancements that have been made to further increase the Fragment Hotspot Maps' domain of utility.

Chapter 2

Development and Theoretical Basis of the Fragment Hotspot Maps Method

2.1 Introduction

This chapter will describe the early work that ultimately formed the rationale behind the Fragment Hotspot Maps method. The first task was to explore the existing hotspot detection methods, comparing how they defined hotspots and what could be learnt about their nature. This information was used to guide how Fragment Hotspot Maps were calculated and validated.

Atomic interaction methods SuperStar [115] and GRID [106] are two well established programs that are able to locate favourable positions for atomic probes on a protein surface, however they would not necessarily result in an environment suitable for ligand binding. GRID places an atomic probe at each point on a 3D grid placed over a protein, and calculates favourable positions for a given probe using force fields. SuperStar uses data from IsoStar [114], a library of molecular contacts in the Cambridge Structural Database (CSD) [119, 120], to give a propensity for a given probe type at each grid point within the cavity. If an interaction between two groups at a certain distance and angle is favourable, it will occur more frequently in the CSD and therefore have a greater propensity in the SuperStar output. These methods are useful, but tend to find too many favourable regions, as calculations do not include solvent, which would normally interact with polar residues.

Consensus site methods Sub-pockets that are found to bind a variety of chemically diverse probe molecules can be referred to as consensus sites. The number of different probes that bind to the site is considered rather than the binding affinity of the probe. Consensus sites can be identified experimentally through Multiple Solvent Crystal Structures (MSCS) [100, 121, 105] and fragment library screening [25], or predicted computationally using simulated annealing of chemical potential (SACP)[122] or FTMap [123–125].

FTMap uses 16 small molecule probes, which are either purely hydrophobic or contain one or two polar functional groups. FTMap ranks its hotspots by counting the number of different types of probe that bind to a given cluster, resulting in a consensus site, reflecting results from experimental multiple solvent crystal structures. Although consensus sites are ranked by their promiscuity, the simplicity of the probes allows many of them to find a single hotspot, making the single polar interaction (if required) and place their carbon atoms in the hydrophobic region surrounding it.

Unhappy water site methods The role of binding site waters is becoming increasingly prominent in structure-based drug discovery [126–132]. Molecular dynamics methods such as WaterMap [130] calculate the thermodynamic properties of hydration sites, identifying "unhappy water" sites. Water-centric methods have been included as a hotspot detection method as unhappy waters are found within hotspots [132]. Using molecular dynamics in explicit water leads to calculation times of approximately 24 hours, but identification of hydration sites leads to finer grain information about the interactions likely to be the cause of the hotspot.

Mixed solvent molecular dynamics A recent review on mixed solvent molecular dynamics (MD) by Ghanakota and Carlson [133] discusses the wide range of methods available [134–142]. The earliest was from Seco, Luque and Barril, who later developed MDMix. MDMix[143] uses three 20 ns MD simulations, with one in the presence of 20% ethanol and another in 20% acetamide. These probes are chosen as they are highly miscible in water, removing the need for artificial potentials to prevent aggregation, as well as containing the three common interaction types: hydrogen bond donor, hydrogen bond acceptor and hydrophobic. They compare their results to GRID[106]. Without explicit solvation, they find that GRID locates too many polar interaction sites, which correspond to favourable water binding locations. In contrast, MDMix's solvent probes are more selective in displacing water molecules that are displaced by ligands.

2.1.1 Limitations of Current Methods

Hotspot detection methods add the most value when little is known about the target. Of the methods described above, FTMap and SACP are the only methods able to detect hotspots from a global search of an *apo* protein structure, whereas the rest are validated against predefined binding sites. These methods instead aim to map the interactions, or hydration sites in the case of WaterMap, within a binding site to show which are likely to be made by the ligand. None of the methods described above aim to do both.

As with all structure-based computational methods, protein flexibility should be considered. Kozakov and colleagues [144] have shown previously that hotspots are less sensitive to conformational change. Looking specifically at protein–protein interaction hotspots, they found that even if substantial conformational change was required for ligand binding, the hotspots were still detectable from the *apo* structure.

Large changes will affect the mapping of interactions within the binding site. A well known example of such a change is the DFG-in to DFG-out conformation change required for type II inhibition of kinases [145]. This is exemplified in figure 2.1, where a type II p38α inhibitor is overlaid with the *apo* binding site. The morpholine occupies the DFG-out pocket, which is not present in the *apo* structure, and any method that uses the *apo* structure would fail to predict these interactions. However, this is not solely a problem for static methods. MD hotspot prediction approaches have insufficient sampling for such a conformational change to occur [146], and they often require restraints on the protein heavy atoms, preventing even small changes from occurring. Furthermore, not all ligand binding pockets exist in the protein crystal structure until they are ligand bound. Prediction of these cryptic pockets is especially difficult, and requires enhanced sampling MD methods [147–151]. Static methods with fast enough calculation times could be used to post process frames from these MD calculations, however FTMap requires 4-24 hours for a calculation [110].

2.1.2 Choosing a Hotspot Definition

Due to the broad usage of the term hotspot in drug discovery, it is important to first give a precise definition in the context of this thesis. The chosen definition will decide how the method is validated, therefore it is important to select one that can be compared to reliable experimental data.

Early work during this project explored the idea that hotspots contribute a disproportionately large amount to the free energy of binding ($\Delta G_{binding}$). FBDD projects that had affinity

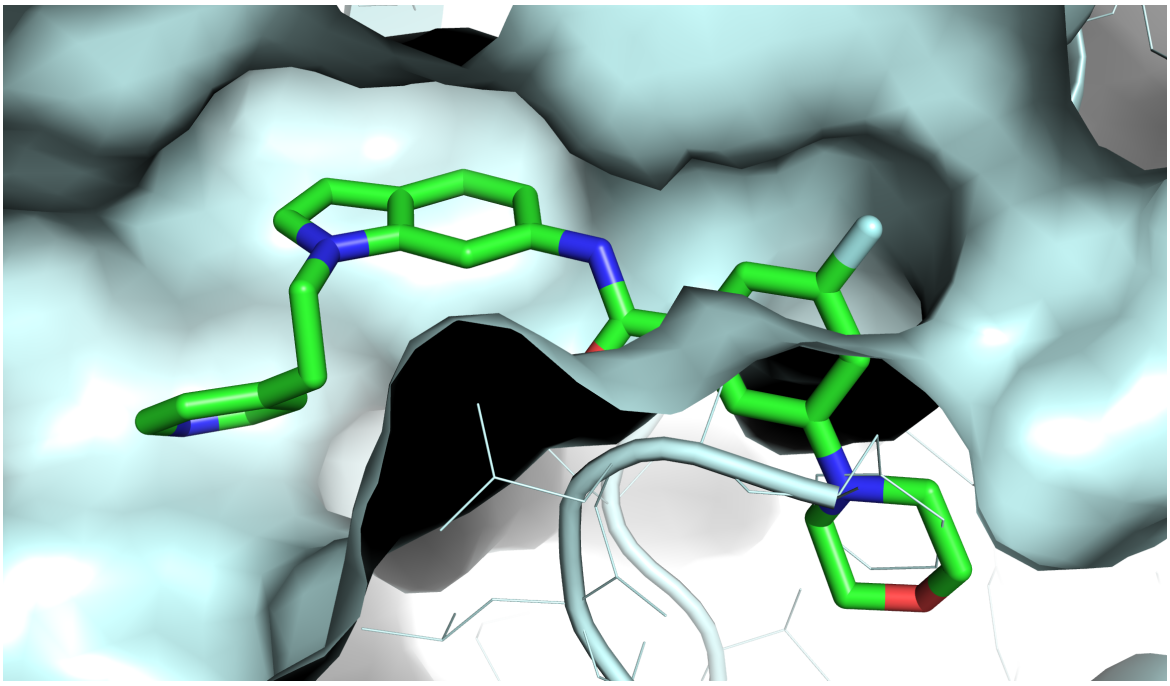


Fig. 2.1 A type II p38α inhibitor (cyan sticks) overlaid with the *apo* binding site (white surface)

data available for the fragment, lead and intermediate steps, in addition to a crystal structure for the lead molecule, were used to create 3D matched molecular pairs (MMPs). An MMP is a pair of molecules with a single transformation between them [152]. This could be a single cut in the case of a terminal group, or a double cut in the case of a core replacement. These were identified using the algorithm implemented by Hussain and colleagues [153] within RDKit.

MMPs are typically used in lead optimisation, aiming to improve target independent properties such as solubility [152]. They are usually not used to predict affinity as this is receptor specific. Addition of a group may have a positive effect for one target, but a negative effect for another. In order to use MMPs to understand activity profiles, they must first be placed in the context of the receptor [154, 155], in this case using the crystal structure of ligand bound protein.

The MMPs between the compounds were filtered such that only additions were included. As a result, the change in $\Delta G_{binding}$ between a MMP ($\Delta\Delta G$s) represents the interactions of the group and are not affected by the loss of another group. The $\Delta\Delta G$s were then divided by the number of heavy atoms to give the group efficiency (GE equation 1.2), which could then be mapped to the centroid of the group using the coordinates from the crystal structure. This

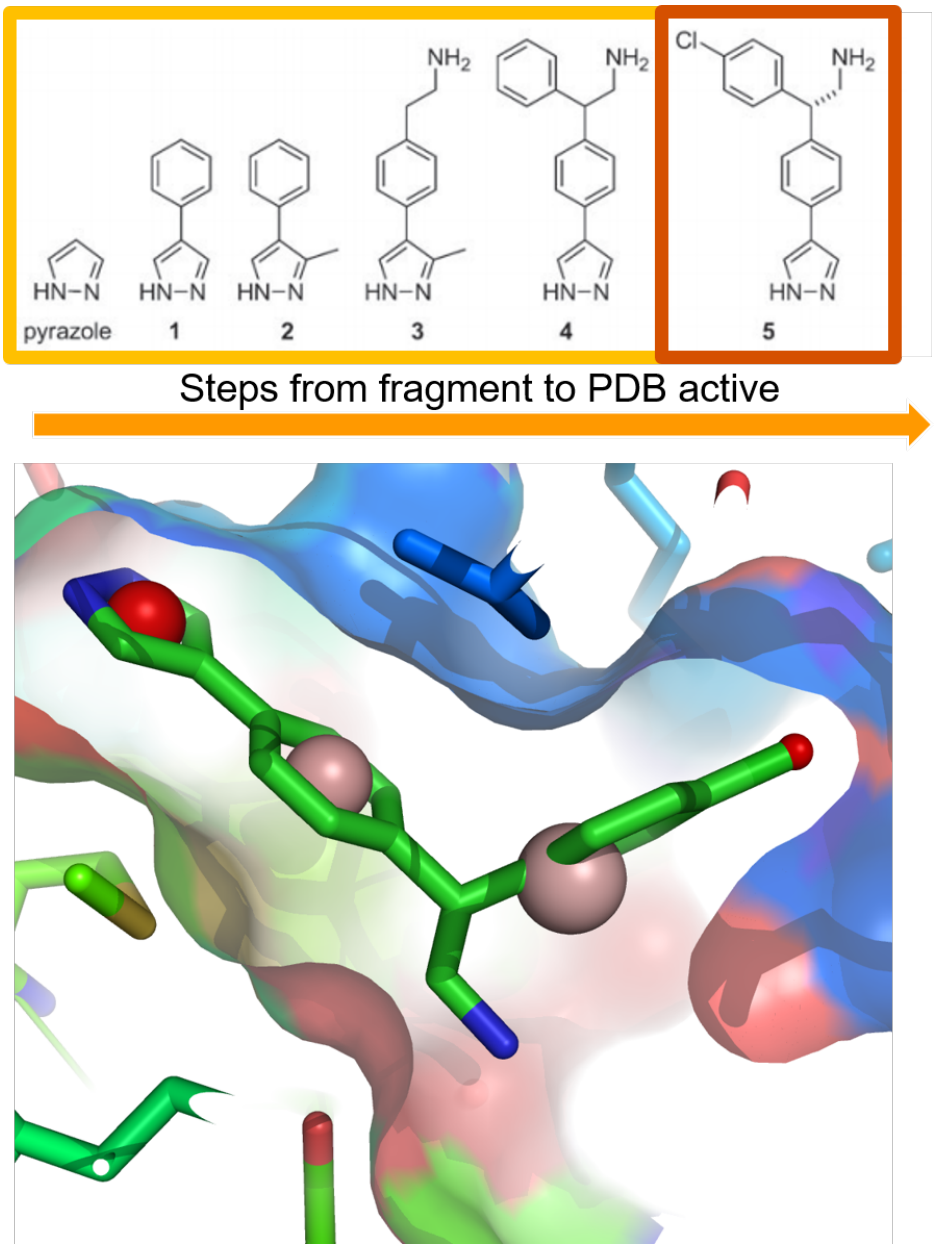


Fig. 2.2 3D Matched Molecular Paris (MMPs) for Protein Kinase B. (Top) 2D structures of compounds used to generate the 3D MMPs, modified from [69]. (Bottom) 3D MMPs are represented by spheres, which are placed at the centroid of the initial pyrazole fragment and each subsequent addition. Darker red denotes a more group efficient moiety and the size of the sphere is related to the number of atoms contributing to the group.

was done for all of the identified MMPs, and displayed as a coloured sphere placed at the centroid for each group.

An example output is shown in figure 2.2, which uses GE data for protein kinase B (PKB) from the publication introducing GE [69]. One important caveat to the data given in this paper is that the loss of rigid body entropy [99] has been accounted for in the calculation of GE for the initial fragment hit.

Other datasets were gathered from both the literature, and systematically from ChEMBL [156]. It soon became apparent that creating a dataset of known hotspots using this approach would be difficult. The method is highly dependent on the amount and quality of data, with few examples of projects able to clearly map the binding site. A requirement of the method was that only additions were made to the molecule, and it was rare to find cases where there was such a linear development for the lead molecule. As a result, the original fragment hit was often not a valid substructure of the lead molecule, leading to the situation in figure 2.3 where the smallest valid molecule made up most of the final structure. The identification of the hotspot through this method relied on inclusion of the rigid body entropy, as shown in figure 2.2. The magnitude of this penalty is large (15-20 $kJmol^{-1}$), and would always make the fragment the most group efficient part of the molecule. Finally, the thermodynamics of protein-ligand binding is affected by many factors on both the macroscopic and microscopic level [157], making it difficult to attribute thermodynamic changes to change in structure.

Moving away from the use of affinity data, attention was turned to X-ray crystallography. Fragment-bound protein crystal structures could be used to define the hotspot within a binding site, however some special considerations would need to be taken into account. The size of a hotspot may not match the size of the fragment bound to it, meaning a given fragment may be able to match the shape and interactions of a hotspot, but also extend outside of it. Secondly, as fragment concentration is very high during a crystallography experiment, it may be that not all fragment binding sites are in fact hotspots. If a fragment is not bound to a hotspot, fragment growth can lead to reorientation in the binding site [158]. Taking all of this into account, for the purpose of this thesis, hotspots will be defined as follows:

> **Hotspots** are the minimum binding site that will bind a fragment, maintaining the fragment binding position once it has been elaborated.

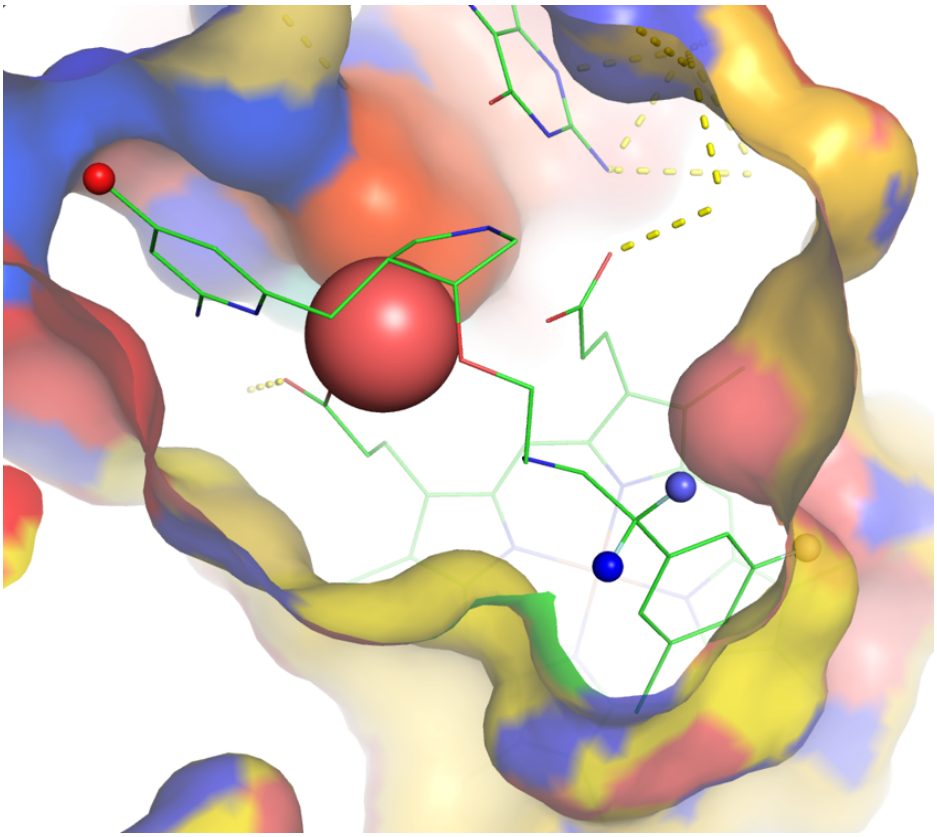


Fig. 2.3 3D Matched Molecular Pairs (MMPs) with insufficient data. Here, the smallest identified compound makes up most of the final compound. Only four single atom additions can be identified, resulting in a large sphere corresponding to the initial hit, with four small spheres for the single atom 3D MMPs.

Table 2.1 Literature descriptions of protein environments that lead to unhappy waters, fragment binding and hotspots

Name	Protein Environment
Unhappy Waters (Young 2007) [130]	A strongly hydrophobic cavity that encloses multiple water molecules or one to three hydrogen bonds with the protein by the ligand, where the remainder of the local environment is hydrophobically enclosed
Fragments (Ichihara 2014) [132]	Fragment hits tend to displace water molecules with notably unfavorable excess entropies— configurationally constrained water molecules. This is likely to be caused by confinement in hydrophobic pockets or a combination of hydrophobic enclosure with hydrogen bonds
Hotspots (Kozakov 2015) [110]	Concave topology combined with a mosaic-like pattern of hydrophobic and polar functionality

2.1.3 Hotspot Environments

To predict the existence of hotspots, it is important to first understand the protein environments that cause them. As discussed previously, fragment binding and high energy hydration (unhappy water) sites occur at hotspots. A literature search for the protein environments found at fragment binding sites, unhappy water sites and hotspots is summarised in table 2.1. All three describe very similar situations:

- Hydrogen bond(s) pointing into an enclosed hydrophobic pocket
- Enclosed hydrophobic environment

The descriptions from Young and Kozakov are relatively brief, however Ichihara *et al.*[132] made this the focus of their paper. They used WaterMap, the subject of Young's paper [130], to compare hydration sites displaced by fragments to those displaced during fragment growing. They found that the most constrained hydration sites were always displaced by fragments, rather than during the process of fragment growing. These were not exclusively hydration sites that had a positive ΔG compared to bulk water, as typically thought of as an unhappy water, but also included those that made a strong polar interaction with the protein. This gave a very negative ΔH, resulting in a favourable change in free energy. Despite the

favourable free energy of hydration, displacement of the water is possible if the ligand is able to replace a geometrically strained water-protein hydrogen bond. This results in a large entropy gain with minimal enthalpy loss. They reasoned that these hotspots are caused by either water interacting with a hydrogen bond located within a confined hydrophobic pocket or simply within a hydrophobic pocket. These environments will result in a reduced number of available orientations of the water molecule, resulting in reduced entropy compared to bulk water.

Water-centric approaches are computationally intensive, with calculations typically taking 24 hours. They are usually used once a binding site has been identified, rather than to predict ligandable pockets. Vukovic and colleagues [159] have developed a method that analyses clusters of high energy hydration sites as a means assessing ligandability, however calculations can take days in addition to the initial 24 hour calculation.

The importance of solvation in the computational prediction of hotspots can be seen from the evolution of computational methods. Early approaches such as MCSS[101] and GRID[106] did not account for solvation, and found hundreds of false-positive minima across the protein surface, in addition to the true-positives found experimentally [103, 105, 100, 121]. Sheldon Dennis and colleagues[160], the group that later developed FTMap, were the first to produce a computational solvent mapping program that was able to reproduce the experimental results. Importantly, their calculations included a desolvation energy term, calculated using a continuum electrostatics model[161, 162]. In the development of FTMap [123], they aimed to improve the speed of their sampling by using a fast fourier fransform (FFT) correlation approach to sample a dense six dimensional translational and rotation grid. This limited their scoring to sums of correlation functions, restricting them to simple energy expressions. The energy expression they developed included terms for van der Waals, electrostatics, a cavity term (to describe contribution of hydrophobic enclosure) and a statistical pairwise potential (to represent other solvation effects). The latter two terms provide a simplified description of how the protein environment affect solvation. Overall, no loss in accuracy was found compared to their original implementation, and a six-fold improvement in speed was achieved.

The approach described in this chapter aims to find the protein environments that cause hotspots and unhappy waters. Buriedness measures and careful selection of probes will be used to find these environments, removing the need for explicit water and MD, resulting in faster calculations.

2.2 Fragment Hotspot Maps Method

2.2.1 Overview

The Fragment Hotspot Maps method was developed in this project to make use of interaction data in the CSD to identify hotspot-yielding environments from a static protein structure. The method requires no prior knowledge of the binding site, performing a global search across the whole protein to locate hotspots. The images in figure 2.4 show the output from the three key stages of the fragment hotspots method, which can be summarised as follows:

Atomic propensity — Mapping the propensities for atomic probes throughout the protein (figure 2.4a)

Buriedness weighting — Weighting the atomic propensities by the grid point's buriedness introduces the required enclosure (figure 2.4b)

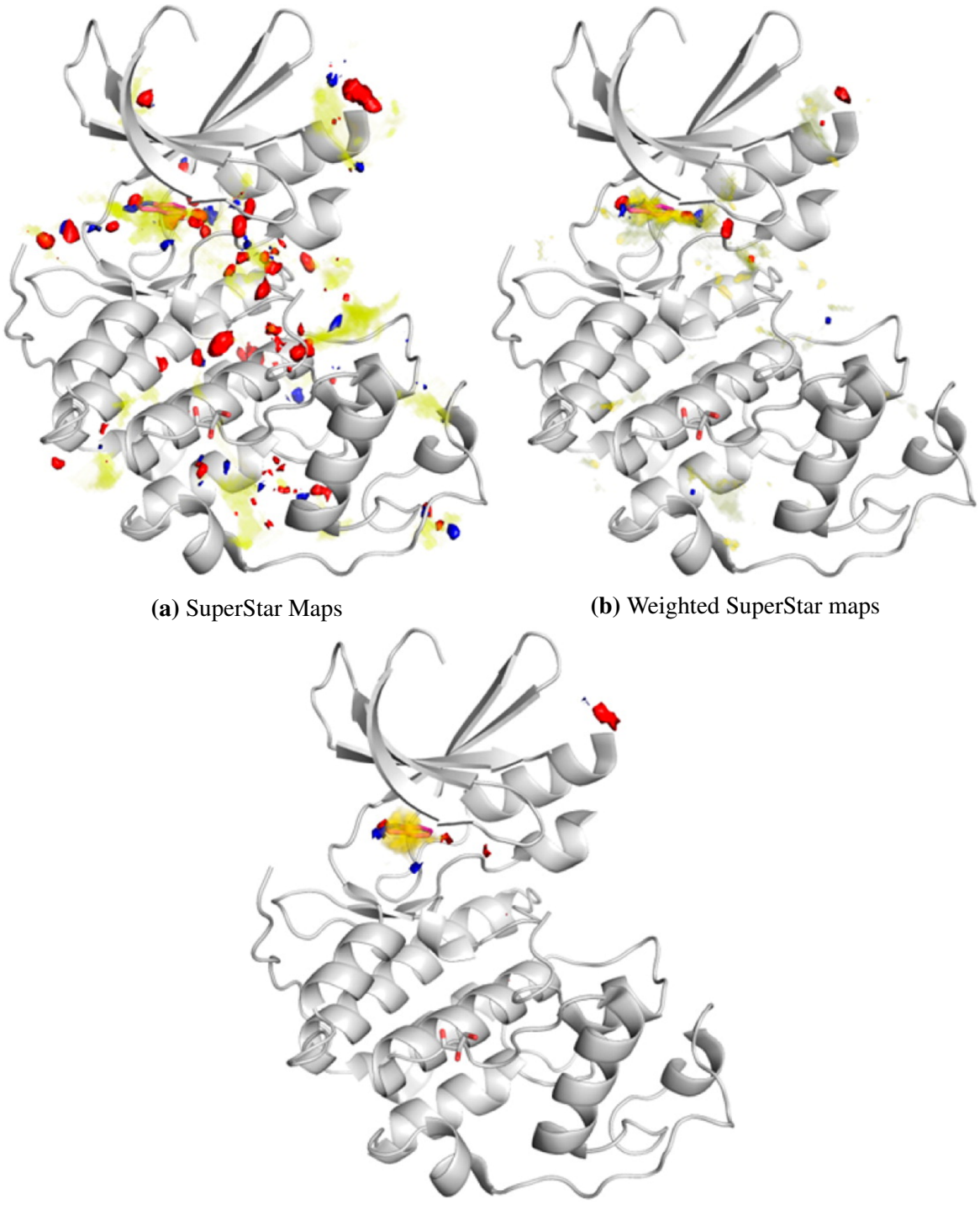
Molecular probe sampling — Final fragment hotspot map output (figure 2.4c). Sampling the weighted propensities with molecular probes has two important effects: eliminate pockets too small for fragment binding, and locate polar interactions found within a hydrophobic environment.

2.2.2 Atomic Propensities

The first step in the fragment hotspot map method is the calculation of atomic propensities. This is done using existing software, SuperStar [115]. SuperStar uses Isostar data [114], a library of intermolecular interactions in the CSD, to map the propensity for a given probe type onto a 0.5 Å grid across the protein. The resulting propensities reflect how many times more likely than random the given probe will be found at that grid point, based on interactions in the CSD.

SuperStar requires the proteins to be protonated, and to have unimportant water or ligand molecules removed. If any water or ligand is left within the protein, it is included in the calculation and treated in the same manner as the protein. This is useful if there is a known bridging water molecule that needs to be included in the calculation.

Normally the binding site would need to be defined prior to the SuperStar calculation; however, no information about the binding site is used in this case. SuperStar uses the



(a) SuperStar Maps

(b) Weighted SuperStar maps

(c) Fragment Hotspot Maps

Fig. 2.4 Output maps at each stage of the Fragment Hotspot Map Calculation. Donor maps are shown in blue, acceptor maps in red and hydrophobe in yellow. A fragment-bound structure (magenta, ligand in sticks and protein hidden) has been aligned to the *apo* structure (white cartoon) for which the maps were calculated, for reference. Taken from Radoux *et al* [1]

LIGSITE [163] algorithm to detect cavities, and in the absence of a starting coordinate or residue from which to grow the cavity, LIGSITE is run on the whole protein. It gives each grid point a buriedness score between zero (completely solvent exposed) and seven (completely buried). SuperStar then provides atomic propensities for cavities that contained grid points with a LIGSITE score of five or above. The three maps shown in figure 2.4a were generated using the SuperStar atomic probes listed below:

Hydrophobic Aromatic CH probe

Donor Uncharged NH probe

Acceptor Carbonyl oxygen probe

To find areas where high interaction propensity coincides with buried pockets, the SuperStar maps are weighted by the LIGSITE score for each grid point. The weighted SuperStar maps (figure 2.4b) begin to highlight the binding site, but still show propensity throughout the protein. To find fragment hotspots, the weighted propensities are sampled with molecular probes.

2.2.3 Sampling with Molecular Probes

As only hydrophobic, donor and acceptor maps are calculated, probes containing either all carbons or carbons with a single donor or acceptor heteroatom are able to sample the maps fully. The probes, shown in figure 2.5, were chosen to reflect hotspot environments. All three probes have the same shape: the polar atoms represent a functional group attached to the ring and toluene is used for the apolar probe. The large but flat rings are selected to find tight hydrophobic environments, with polar interactions at the deepest part for the polar probes. The probes may be too large to sample very small pockets accessible to alkyl chains but were chosen as they resulted in fewer false positives when performing a global search. Smaller probes could be used to give a deeper exploration of pockets highlighted by the default probes. The bond orders of the probes are ignored, and it is just the atom types that are used to assign a score from one of the three weighted SuperStar maps.

The probes undergo 200 rotations, which are uniformly distributed on the surface of a sphere then translated such that they are centred on the heteroatom for the polar probes or the methyl group of the toluene probe. All rotations of a probe are placed with their central atom on the top 400 scoring grid points. Since the publication of the method's validation

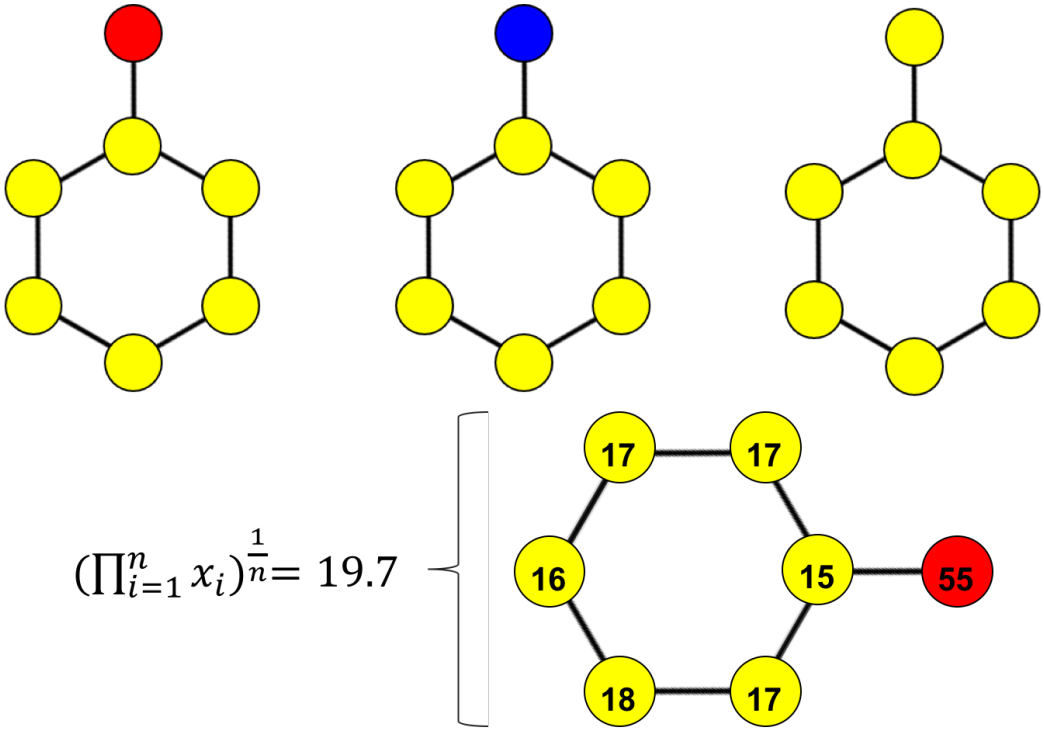


Fig. 2.5 Three molecular probes used to sample the weighted atomic propensity maps. The bonds of the probes are ignored and the atom types depend on the SuperStar probe used, in this case all yellow atoms sample the aromatic CH propensity, the red atom samples the carbonyl oxygen propensity, and the blue atom samples the uncharged NH_2 propensity. The probe at the bottom shows an example of how the probe score is calculated.

[1], sampling has been modified to handle proteins of different sizes better. The most recent method uses 1000 rotations and translates the probes to all grid points with a weighted propensity greater than 17 (by default).

Calculation of the atomic propensities prior to sampling drastically reduces the number of translations required. The total number of poses is in the range of hundreds of thousands to millions, whereas FTMap is required to sample billions of poses[123].

For each pose, the atomic propensities are assigned to each atom from their corresponding map. Any atom that clashes with the protein has a score of zero and the pose is skipped; all remaining poses are assigned a score calculated by taking the geometric mean of their atomic scores. Three 0.5 Å grids, one for apolar, donor, or acceptor atoms, are placed over the protein. Each grid point that contains a probe atom is set to the score of the probe, not including the carbon atoms for the polar probes. If multiple probes place atoms in the same grid point, the highest score is used, giving the final output shown in figure 2.4c.

2.2.4 Fragment Hotspot Map Output

The resulting fragment hotspot maps are output as three grid files for each of the molecular probes. The grids contain the scores of the molecular probes at each grid point, and require an isosurface contour at a given score in order to be visualised. An example output is given for CDK2 in figure 2.6, showing contours at 0, 13 and 18. A contour of 0 shows every grid point that has been sampled by a probe, regardless of the score. At 13, the maps cover all of the ATP binding site as well as several other pockets around the protein. Finally, a contour of 18 shows the highest scoring region of not only the protein, but also within the binding site itself, picking out the interactions made by the fragment.

The scores themselves do not represent any measurable experimental value, but instead describe how well a particular interaction or region resembles a hotspot yielding environment. A high scoring region of the map represents a strong polar interaction (if applicable), found in a highly buried and hydrophobic environment. Given the chosen definition of hotspot, the highest scoring regions of the Fragment Hotspot Maps should match the fragment binding site, with the high scoring polar interactions matching those made by the fragment. This is demonstrated to be the case in figure 2.6, however this has been extended to a wider validation set, covered in the next chapter.

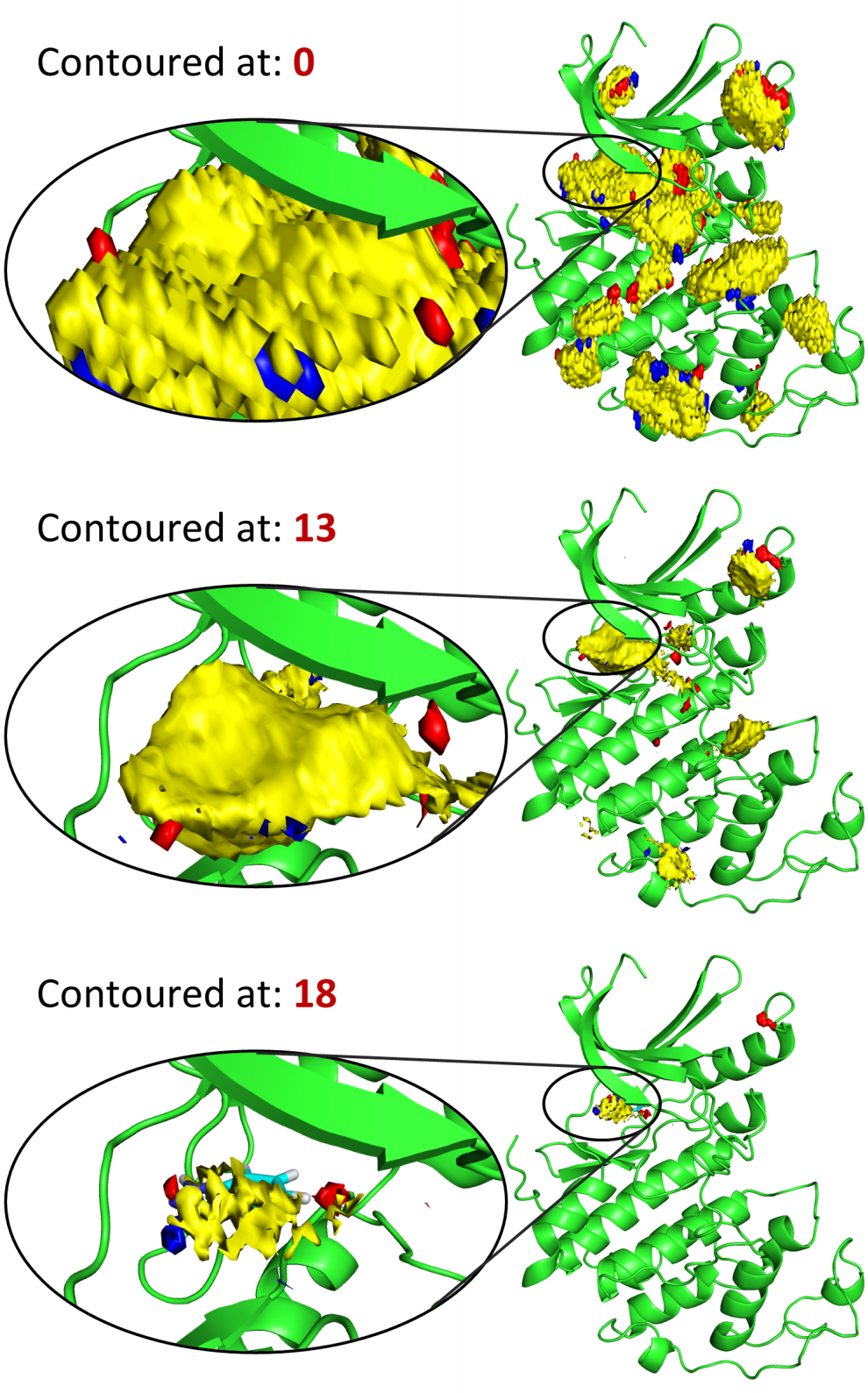


Fig. 2.6 CDK2 with Fragment Hotspot Maps at different score contours. The maps were calculated for an *apo* crystal structure of CDK2, and a fragment bound structure (cyan) aligned to provide a reference

2.3 Conclusion

This chapter has discussed the key characteristics of small molecule binding hotspots, and outlined the Fragment Hotspot Method, which aims to identify those environments. Literature describing unhappy waters, fragment binding sites and hotspots each identified that the protein environment at these sites provided hydrophobic enclosure, often with hydrogen bonds trapping water molecules. The first computational method [160] capable of reproducing the experimental results from MSCS, FTMap, improved upon existing methods by including a desolvation term.

For Fragment Hotspot Maps, the effect of solvation is considered implicitly through the inclusion of buriedness and selection of probes that identify environments shown to yield unhappy waters [130], fragment binding [132] and hotspots [110]. The atomic interaction propensities calculated by SuperStar are weighted by this buriedness term, and then sampled by three molecular probes. The probe scores are used to generate the output Fragment Hotspot Maps, with calculations taking around 10 minutes on three processors.

Fragment Hotspot Maps give a continuum of scores that can be contoured to give different levels of description. Lower contours can describe whole pockets or warm areas, while increasing the score contour will identify the hottest part of the binding site. The chosen hotspot definition stated that fragments should bind to hotspots without changing their binding pose upon elaboration. This defines how the method should be validated, which will be discussed in the next chapter.

Chapter 3

Validation of the Fragment Hotspot Maps Method

3.1 Validation Method

It is assumed that at least part of a fragment must interact with a hotspot in order to bind efficiently enough to be both detectable by X-ray crystallography and to remain in place upon elaboration. Fragment binding positions were therefore used as a standard for hotspot prediction, with the understanding that it is possible that only part of the fragment will be located within the hotspot. To discriminate between hotspots and the rest of the ligand binding site, atoms from lead-like molecules were also examined and compared to the fragments.

The dataset collated by Ichihara *et al.* [132] contains crystal structures from fragment-based drug design projects, where lead molecules developed from the fragment hit retain the fragment binding position. Affinity data were available for each of the fragments and leads. Here, this data set of fragment–lead pairs was extended further to include *apo* structures, on which all calculations were performed, removing bias toward the binding site.

Protonated structures were retrieved from the Protoss server [164], which also protonates the ligand using the context of the binding site where all waters and small molecules have been removed. For each protein, the protonation was checked manually and then the fragment and lead bound structures were globally aligned with the *apo* structure. Only the fragment binding monomer was used.

Maps were created for all *apo* structures in the data set and scores assigned to both the fragment and lead atoms. Each atom in the ligand was categorized as either hydrophobic,

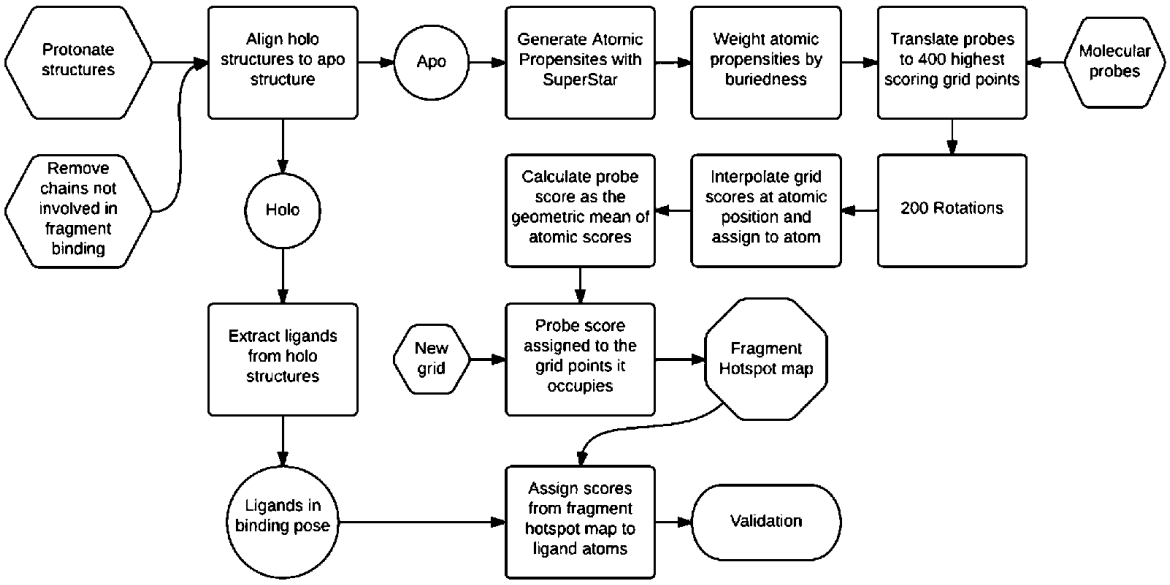


Fig. 3.1 Complete work flow for validation and calculation. Taken from Radoux *et al* [1]

donor or acceptor, and its score was read from the corresponding map. Sometimes the ligand atoms were slightly displaced from a hotspot due to the alignment of the proteins and errors in the crystal structure. To accommodate for this, the highest scoring grid point within two grid points was assigned to the atom. Atoms that were in the maximum common substructure match between the fragment and the lead molecules were assigned as “fragment atoms”, with the remaining atoms assigned as “lead atoms”. The complete work flow for the validation is summarised in figure 3.1.

3.2 Results

In addition to being able to highlight the fragment-binding site, the highest scoring interactions predicted from the apo structure are often those made by the fragment; moderate scoring interactions are picked up by the lead molecule. An example can be seen in figure 3.2, which shows HSP90 with a lead molecule developed by fragment linking. The portion circled in blue coincides with the more potent of the two fragments and can be seen to occupy the highest scoring region of the map (figure 3.2). Only one of the two acceptors is predicted because the fragment binds to HSP90 via bridging water molecules, which were excluded for the purpose of the validation. If a water molecule is known or predicted by other software such as WaterMap [130] to be important for binding it can be included in the calculation.

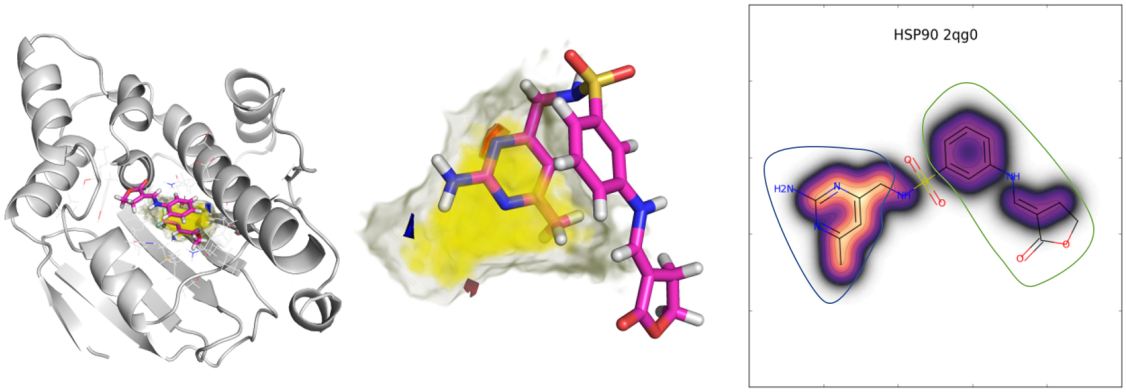


Fig. 3.2 (left) HSP90 with a lead molecule (magenta) and apolar (yellow to grey volume), donor (blue surface), and acceptor (red surface) hotspot maps. (center) A closer look at the ligand with the protein removed. (right) 2D schematic showing how the scores are distributed throughout the lead molecule, with the primary fragment encircled in blue and the second in green. Scores are assigned based on their atom type (e.g., acceptor nitrogen read from the acceptor map). Bright-yellow regions indicate scores >17, purple indicates scores in the range 14–17, and scores lower than 14 are not highlighted. The use of 17 or 14 is discussed below.

For HSP90, inclusion of the relevant water molecules allows identification the remaining interactions of the fragment. The second fragment is shown to have a much lower score. However, the crystal structure with both fragments bound shows it to stack on top of the first fragment, likely contributing to its binding.

The results for all fragment and lead protein complexes are summarised in table 3.1 and figure 3.3. In all cases, the average fragment atom score is greater than the average lead atom score, and in most cases the highest scoring fragment atom was in the the 99th percentile or greater of map scores.

The NS5 RNA polymerase fragment-binding site was the lowest scoring of the data set. The highest scoring atom of the fragment only had a score of 8.6, as the fragment bound to a moderately scoring region known as the thumb site, away from the large highly scoring catalytic region described as the palm([165]). The fragment had the lowest experimentally determined affinity out of the data set, in the mM range, and inspection of the electron density showed that the fragment was poorly resolved. The temperature factors (B-factors) of the fragment atoms ranged from 32 to 42 compared to the surrounding residue atoms, which ranged from 12 to 24. This suggests that pockets should not be thought of as “hot or not”, but rather a continuum, where moderate scoring regions are able to bind fragments, albeit very weakly.

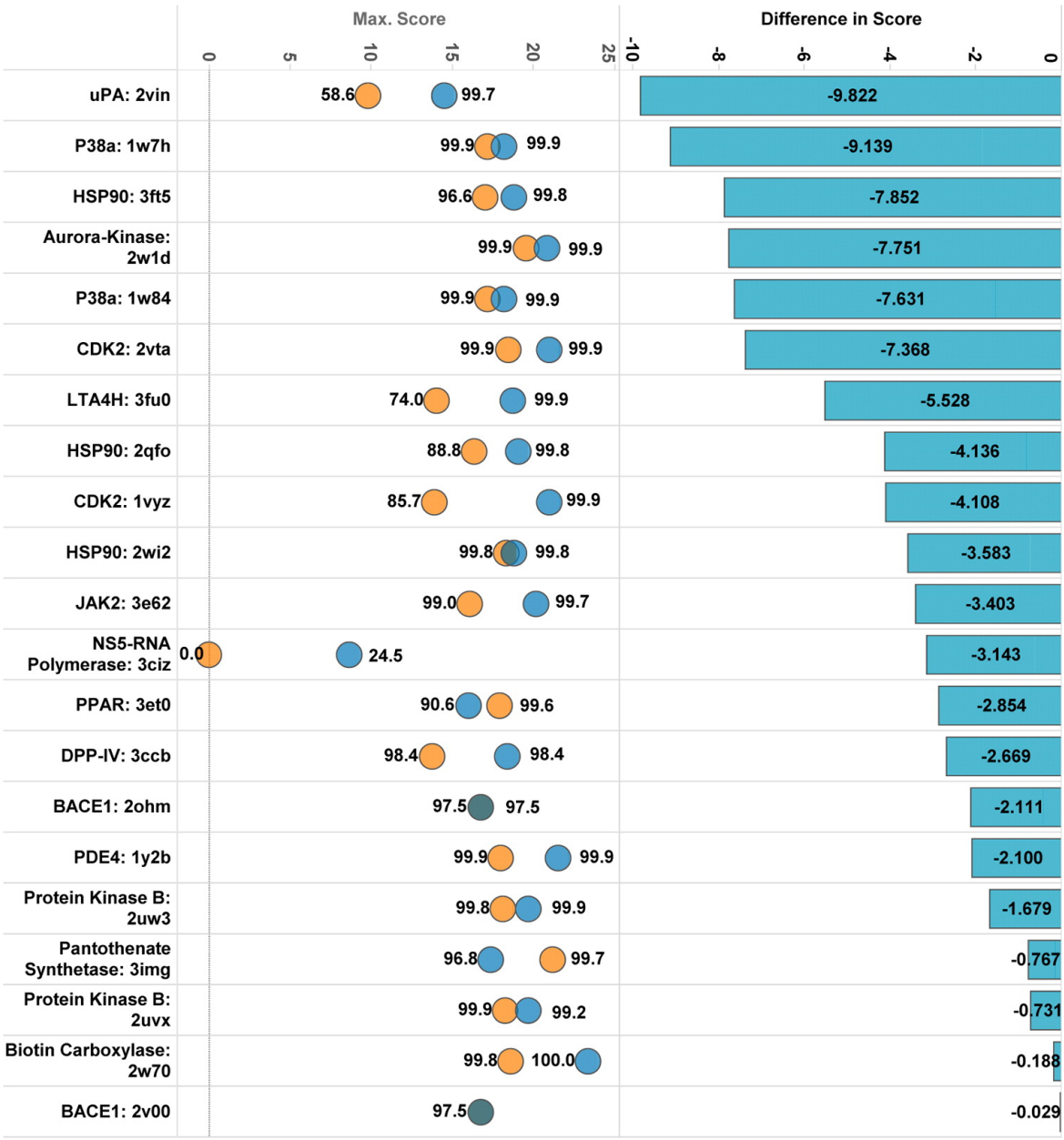
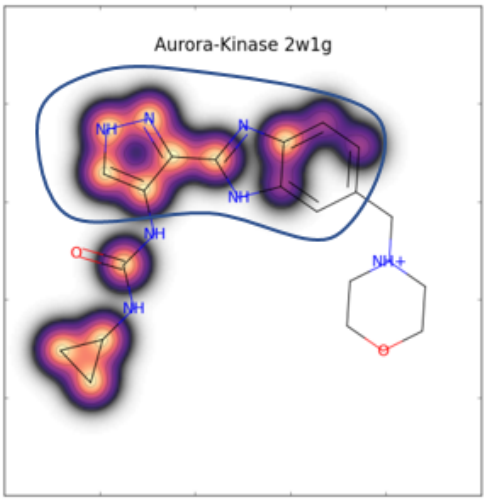
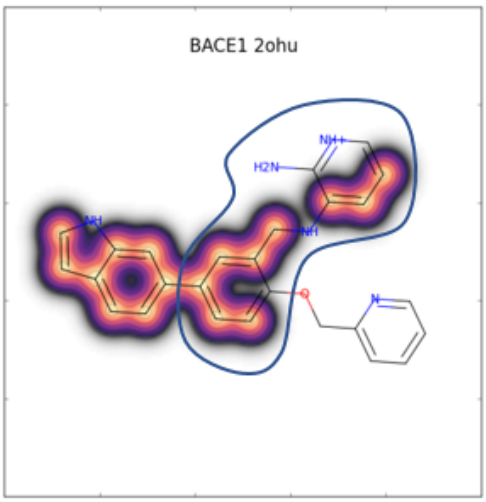
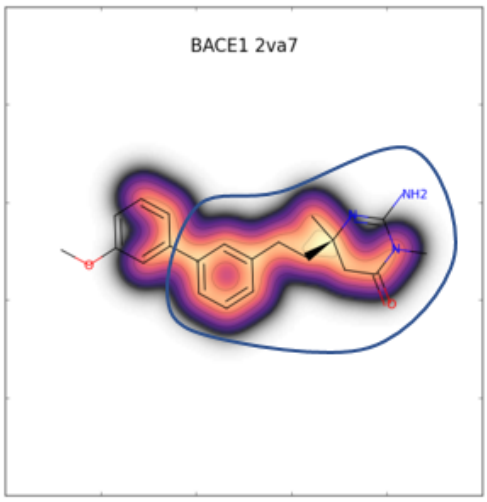
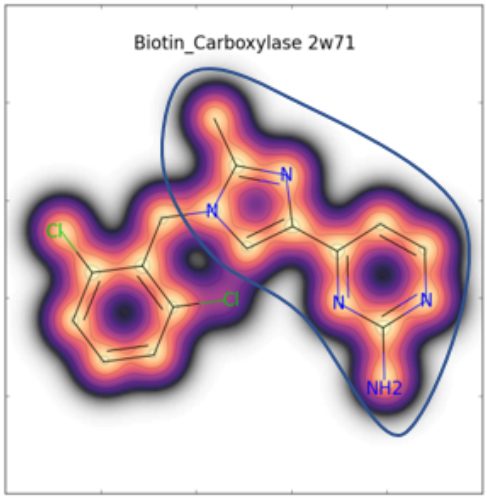
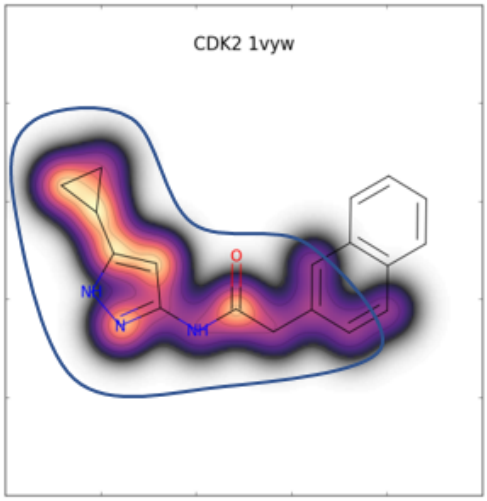


Fig. 3.3 Overview of fragment and lead scores. (left) The highest scoring fragment atom (blue) and lead atom (green) for each fragment-lead pair, labeled by their percentage ranking compared to all grid points with a score greater than 0. At least part of the fragment is expected to interact with a hotspot, therefore the highest scoring atom is used to determine whether the ligand is interacting with a hotspot. Most fragments had their highest scoring atom in the top 1% of scoring grid points. (right) Bar graph showing the (average lead atom score) – (average fragment atom score) for each fragment–lead pair. In all cases, the fragment scores more highly than the lead.

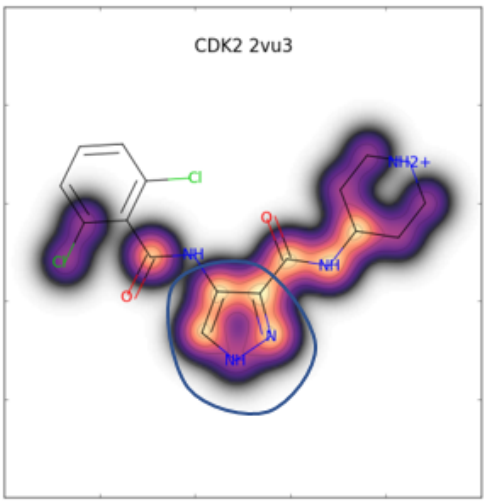
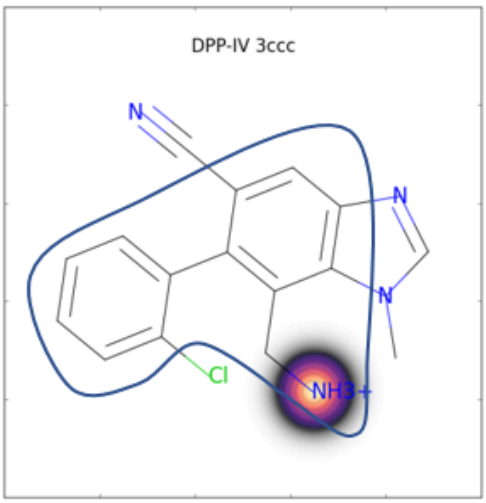
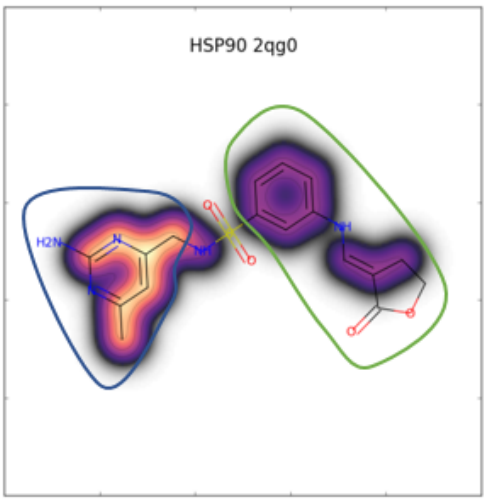
Table 3.1 Overview of datasets and results. The 2D structures of the leads have been mapped with the atomic scores calculated from Fragment Hotspot Maps. Bright-yellow regions indicate scores >17, purple indicates scores in the range 14–17, and scores lower than 14 are not highlighted. The part of the molecule corresponding to the fragment is outlined in blue, and secondary fragments or small molecules present in the fragment crystal structure that were later incorporated into the lead are outlined in orange. Each image is titled by the protein name and accompanied by related PDB codes, and the ligand efficiencies (LE) of the fragment (Frag LE) and the lead (Lead LE). PDB codes labelled with * are structures with the natural substrate (unrelated small molecule inhibitor in the case of JAK2), as an *apo* structure was not available.

	Fragment PDB code: 2W1D Lead PDB code: 2W1G *Apo* PDB code: 4J8N Frag LE: 0.60 Lead LE: 0.43
	Fragment PDB code: 2OHM Lead PDB code: 2OHU *Apo* PDB code: 1W50 Frag LE: 0.33 Lead LE: 0.24

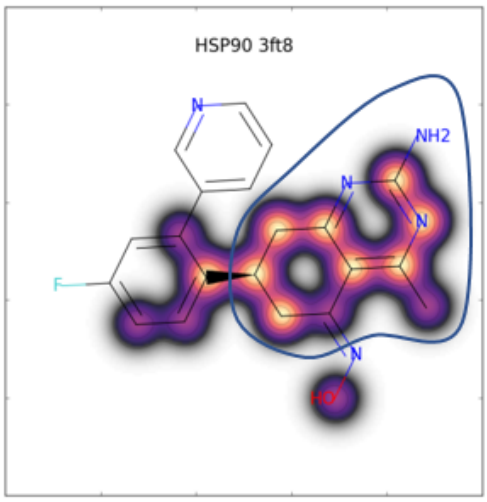
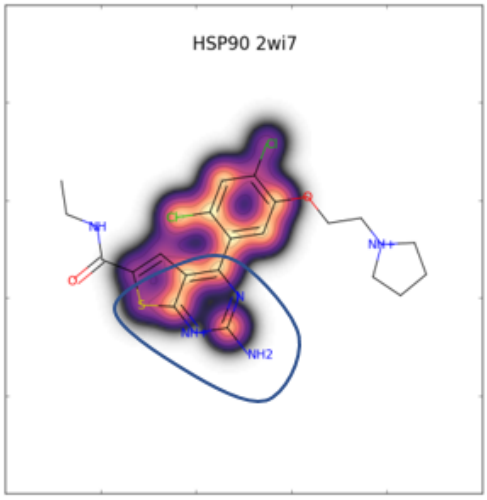
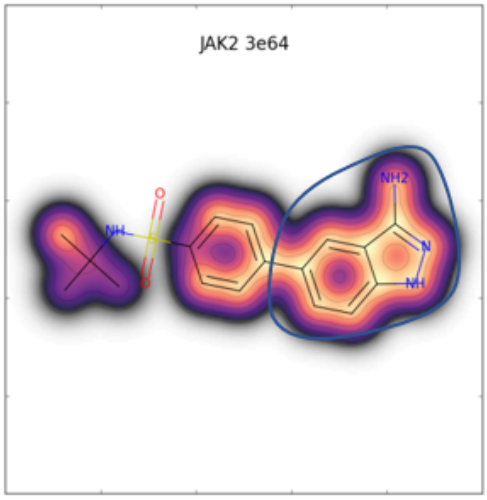
Continues on next page

	Fragment PDB code: 2V00 Lead PDB code: 2VA7 *Apo* PDB code: 1W50 Frag LE: 0.32 Lead LE: 0.36
	Fragment PDB code: 2W70 Lead PDB code: 2W71 *Apo* PDB code: 2J9G* Frag LE: 0.53 Lead LE: 0.41
	Fragment PDB code: 1VYZ Lead PDB code: 1VYW *Apo* PDB code: 1HCL Frag LE: 0.58 Lead LE: 0.41

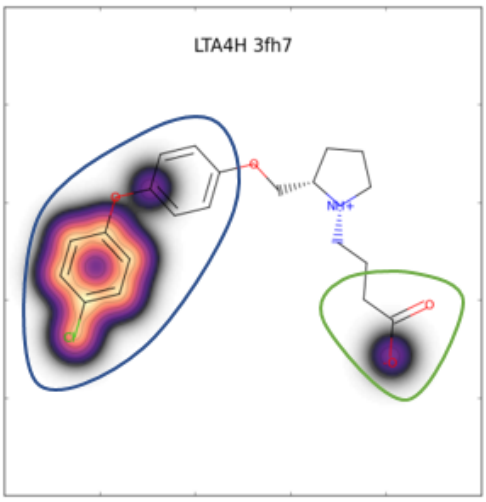
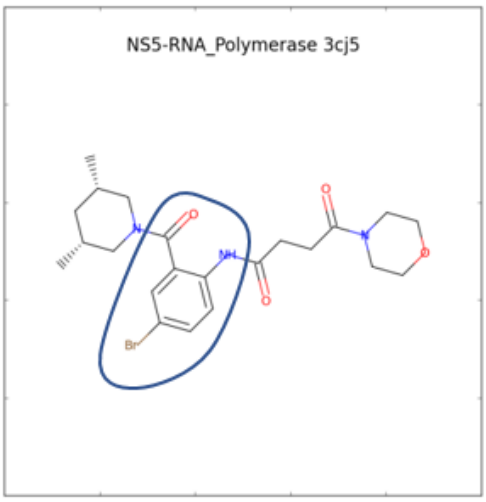
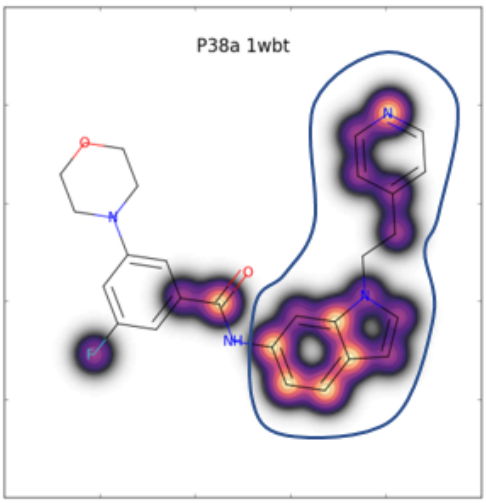
Continues on next page

	Fragment PDB code: 2VTA Lead PDB code: 2VU3 *Apo* PDB code: 1HCL Frag LE: 0.54 Lead LE: 0.47
	Fragment PDB code: 3CCB Lead PDB code: 3CCC *Apo* PDB code: 1J2E Frag LE: 0.45 Lead LE: 0.54
	Fragment PDB code: 2QFO Lead PDB code: 2QG0 *Apo* PDB code: 1YES Frag LE: 0.55 Lead LE: 0.30

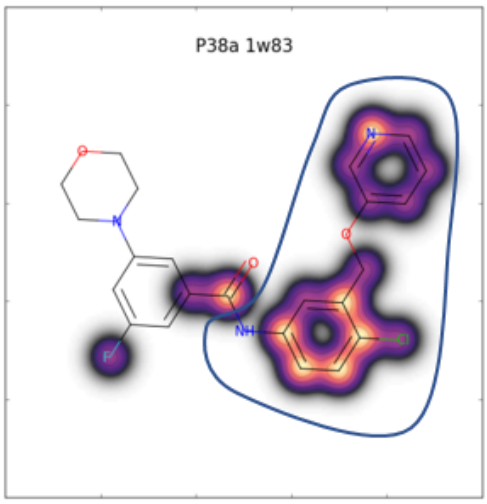
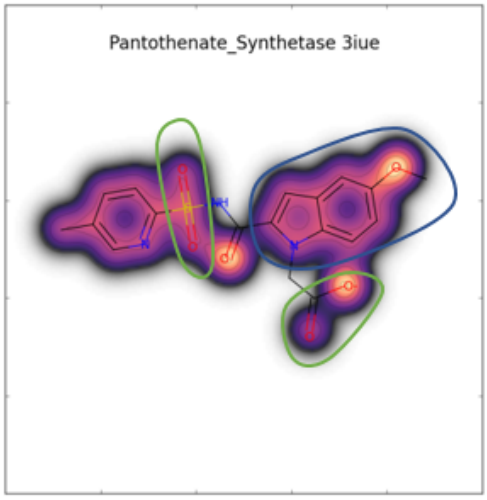
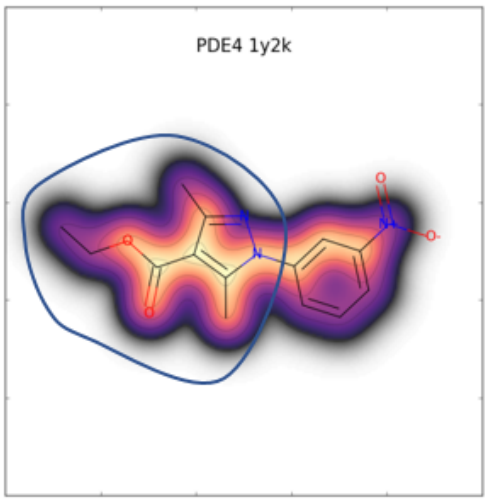
Continues on next page

	Fragment PDB code: 3FT5 Lead PDB code: 3FT8 *Apo* PDB code: 1YES Frag LE: 0.56 Lead LE: 0.39
	Fragment PDB code: 2WI2 Lead PDB code: 2WI7 *Apo* PDB code: 1YES Frag LE: 0.48 Lead LE: 0.33
	Fragment PDB code: 3E62 Lead PDB code: 3E64 *Apo* PDB code: 4ZIM* Frag LE: 0.56 Lead LE: 0.41

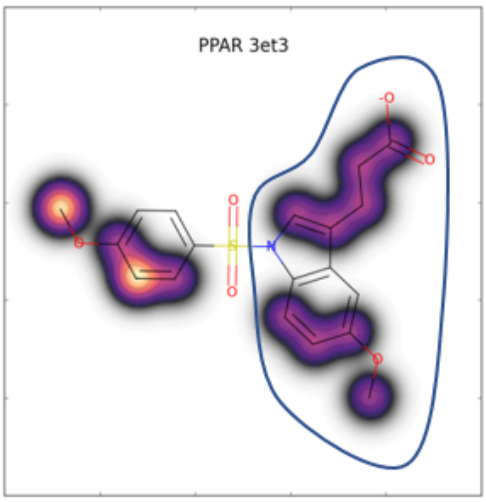
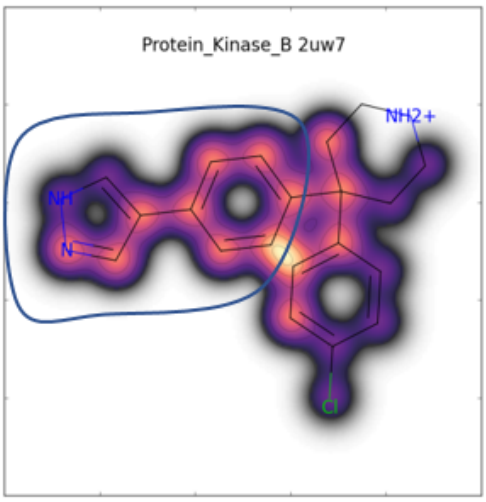
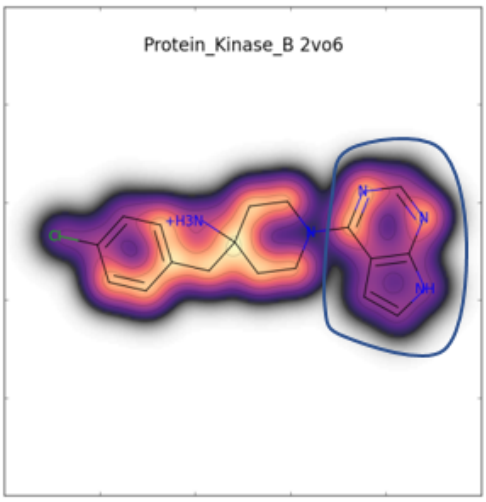
Continues on next page

	Fragment PDB code: 3FU0 Lead PDB code: 3FH7 *Apo* PDB code: 3B7S* Frag LE: 0.21 Lead LE: 0.39
	Fragment PDB code: 3CIZ Lead PDB code: 3CJ5 *Apo* PDB code: 3MWV Frag LE: 0.25 Lead LE: 0.31
	Fragment PDB code: 1W84 Lead PDB code: 1WBT *Apo* PDB code: 1WFC Frag LE: 0.37 Lead LE: 0.32

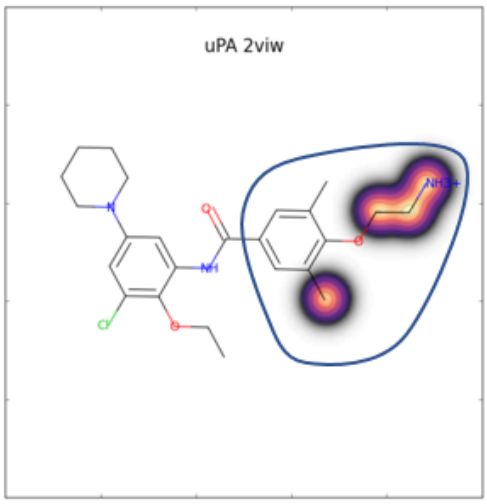
Continues on next page

	Fragment PDB code: 1W7H Lead PDB code: 1W83 *Apo* PDB code: 1WFC Frag LE: 0.28 Lead LE: 0.27
	Fragment PDB code: 3IMG Lead PDB code: 3IUE *Apo* PDB code: 3COV Frag LE: 0.38 Lead LE: 0.29
	Fragment PDB code: 1Y2B Lead PDB code: 1Y2K *Apo* PDB code: 3SL3 Frag LE: 0.48 Lead LE: 0.51

Continues on next page

	Fragment PDB code: 3ET0 Lead PDB code: 3ET3 *Apo* PDB code: 1PRG Frag LE: 0.33 Lead LE: 0.31
	Fragment PDB code: 2UW3 Lead PDB code: 2UW7 *Apo* PDB code: 4C33 Frag LE: 0.48 Lead LE: 0.45
	Fragment PDB code: 2UVX Lead PDB code: 2VO6 *Apo* PDB code: 4C33 Frag LE: 0.59 Lead LE: 0.48

Continues on next page



Fragment PDB code: 2VIN
Lead PDB code: 2VIW
Apo PDB code: 3OY5
Frag LE: 0.32
Lead LE: 0.32

The method aims to locate the interactions that drive fragment binding starting from a global search of the protein. It is therefore important that the method avoids returning false positive binding sites. Here it was assumed that the experimental binding sites from the data sets are the only true positive binding sites, and any other sites detected were treated as false positives.

During the calculation of the atomic propensities, the cavity detection process disregards much of the protein surface. After sampling the atomic propensities with the molecular probes and generating maps based on the highest scoring poses of each probe, the majority of probes could only be placed within a small subset of cavities. To check whether the fragments were found in the highest scoring regions, the atomic scores were compared only to grid points that had at least one probe atom placed there. For each atom in the data set, its score was ranked against all qualifying grid points of the protein that it was calculated from, and was represented as a percentage rank.

Fragments were found to rank more highly than the lead atoms, with a median rank of 97% compared to 72% for the lead atoms (those outside the fragment core). This demonstrates that the Fragment Hotspot Maps are not simply locating ligand-binding sites but are picking out the hotspot within those sites.

To aid with visualizing the output and assessing whether a hotspot is present, a cut-off for a predicted hotspot was calculated. As it is possible that the fragment is larger than the hotspot it binds to, the upper quartile of atomic scores for each fragment was used. The

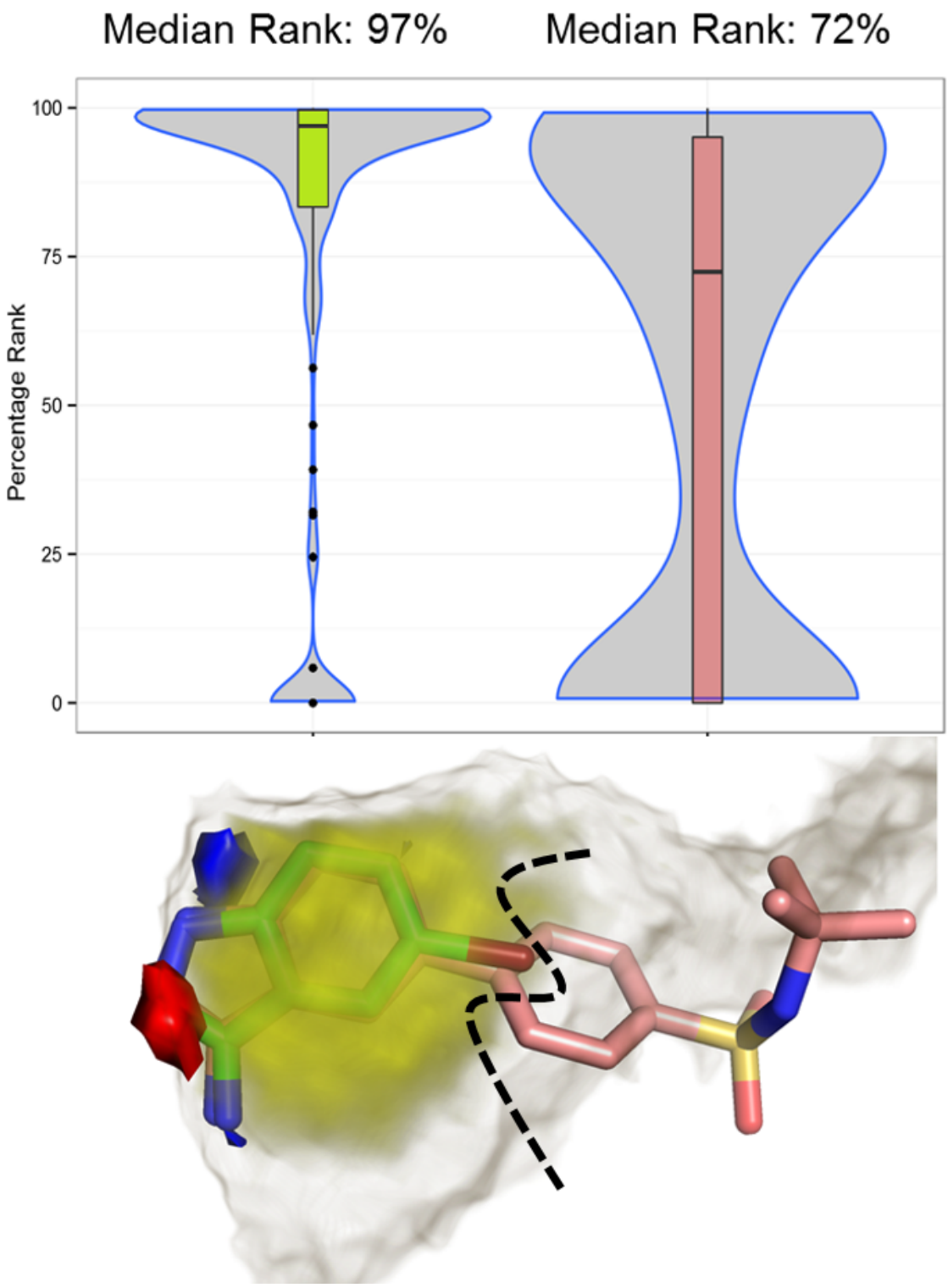


Fig. 3.4 Box and violin plots showing the percentage rank for fragment (green) and lead (pink) atoms. An example fragment-lead pair within the Fragment Hotspot Maps is shown below the plot in the same colours as the plot. The dotted line divides those atoms that contribute to the fragment from those that contribute to the lead.

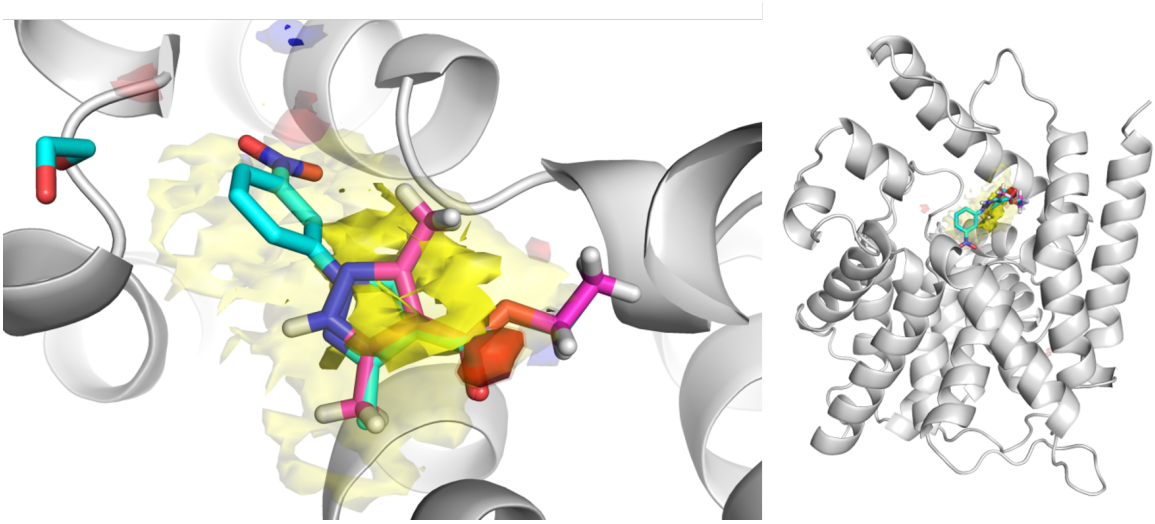
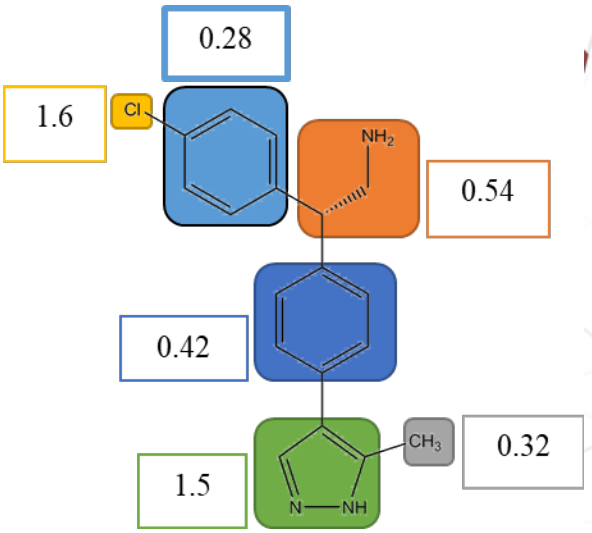


Fig. 3.5 (left) PDE4 with a fragment in magenta and a lead molecule in cyan. The maps have been contoured once at 14 (semitransparent) and again at 17 (almost opaque). Many of the fragment atoms coincide with high scoring apolar (yellow) and acceptor (red) regions, suggesting these interactions drive fragment binding. The cyan lead atoms and some of the fragment atoms extend into regions of the pocket that only make the lower contouring level. The maps do not predict the NH of the fragment, as it is facing the solvent. (right) At this contouring level, only the binding site contains any surfaces, despite starting from a global search.

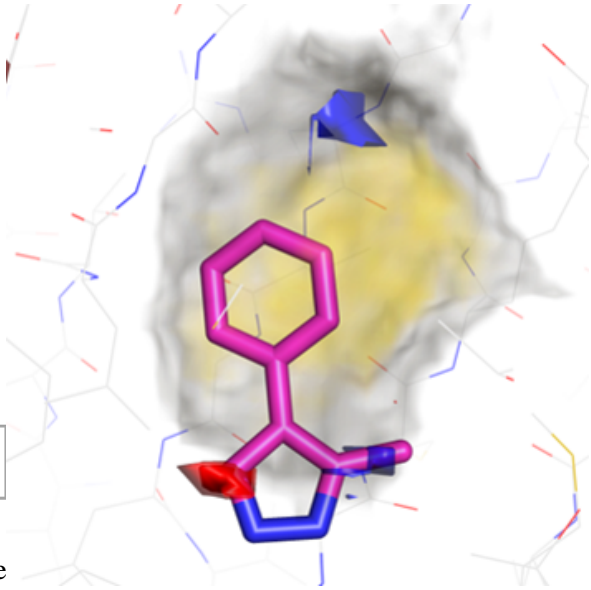
median of these values across the data set was 17, and the lowest was 14. Contouring at these two levels allows visualization of not only where on the protein ligands are likely to bind but also where within that pocket the fragment will bind and which interactions will drive binding. This can be seen in figure 3.5, where areas of acceptor and apolar propensity >17 suggest the interactions leading to fragment binding, with areas >14 matching lead atoms and remaining fragment atoms. Only the binding site of PDE4 contains maps scoring above these contour levels; therefore this information does not come at a cost of being unable to identify the binding site from a global search.

3.2.1 Protein Kinase B

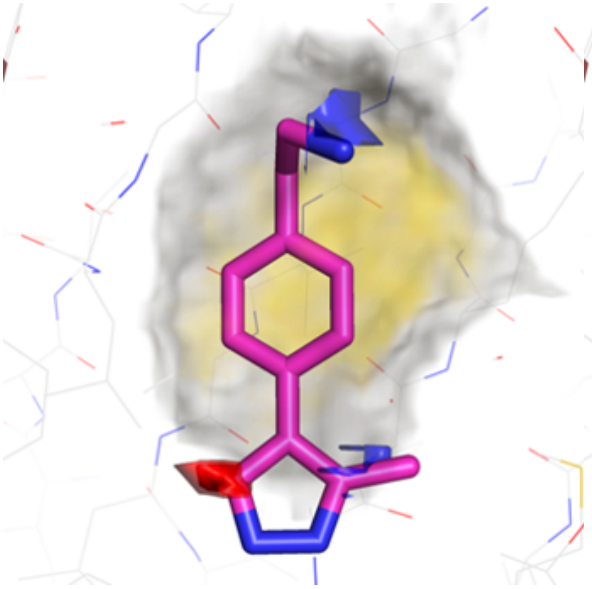
Verdonk *et al.* [69] used a fragment growing approach to design inhibitors of protein kinase B (PKB), and performed a Free–Wilson analysis to provide group contributions to binding. They used GE, defined in equation 1.2, to evaluate whether a group increased potency sufficiently to justify the number of heavy atoms it contained.



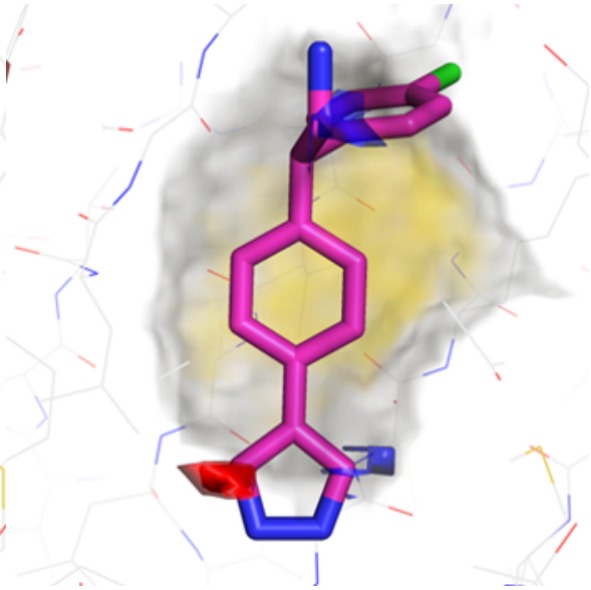
(a) Inhibitor of protein kinase B labelled with the group efficiency of each highlighted group [69].



(b) Fragment hit.



(c) Intermediate structure during fragment growth.



(d) Full compound.

Fig. 3.6 Breakdown of PKB's GE. Hydrophobic map in dark-grey to yellow shows moderate to high scoring regions. Donor hotspots are shown as a blue surface, and acceptor hotspots are shown as a red surface. The atoms seem displaced from the maps as the global alignment of the proteins did not manage to align the binding site well.

Their results are summarised in figure 3.6. The pyrazole is estimated to have a group efficiency of 1.5, second only to the single-atom chloro group. Both the acceptor and the donor interactions are predicted in the Fragment Hotspot Map, and the carbons are located within the main hydrophobic hotspot. Comprising only five atoms, this small fragment binds efficiently, as expected.

The first phenyl group does not make any specific interactions but is located mostly within the main hydrophobic hotspot, reflected in a GE of 0.42. The methyl group has one of the lowest group efficiencies of the groups, and was ultimately removed from the molecule. From the maps, it can be seen to extend slightly outside of the main hotspot.

Addition of -EtNH_2 to the phenyl ring yields one of the more group efficient additions to the molecule. This is easily rationalized from the map in figure 3.6c, as the primary amine occupies a region of high scoring donor propensity.

The second phenyl group is given a group efficiency of 0.28, which is the lowest of all groups. This could be an underestimate, as addition of the phenyl group prevents the primary amine from making the interaction it made previously, as can be seen in figure 3.6d. This is one example out of many where addition of a group does not make a simple additive increase in potency [166].

If the fragment binding site is known, it no longer makes sense to do a global search of the protein unless additional binding sites are of interest. Instead, the calculation can be run with the fragment included and the binding site defined to find nearby "warm spots". This increases the speed of the calculation and limits the information to the cavities of interest to the medicinal chemistry program. The run time for this calculation is only a few minutes on a single processor.

The result of this calculation is shown in figure 3.7, where it is much more obvious which direction the fragment should be grown in order to make the most efficient addition to the molecule, with part of the phenyl group and the chloro group placed in highly scoring areas. As the chloro group is a single atom placed in a highly scoring region, it is understandable why it is the most group efficient addition. The chloro atom also affects the electronics of the phenyl ring, and could again be an example where addition of a group does not lead to simple additive increase in potency [166].

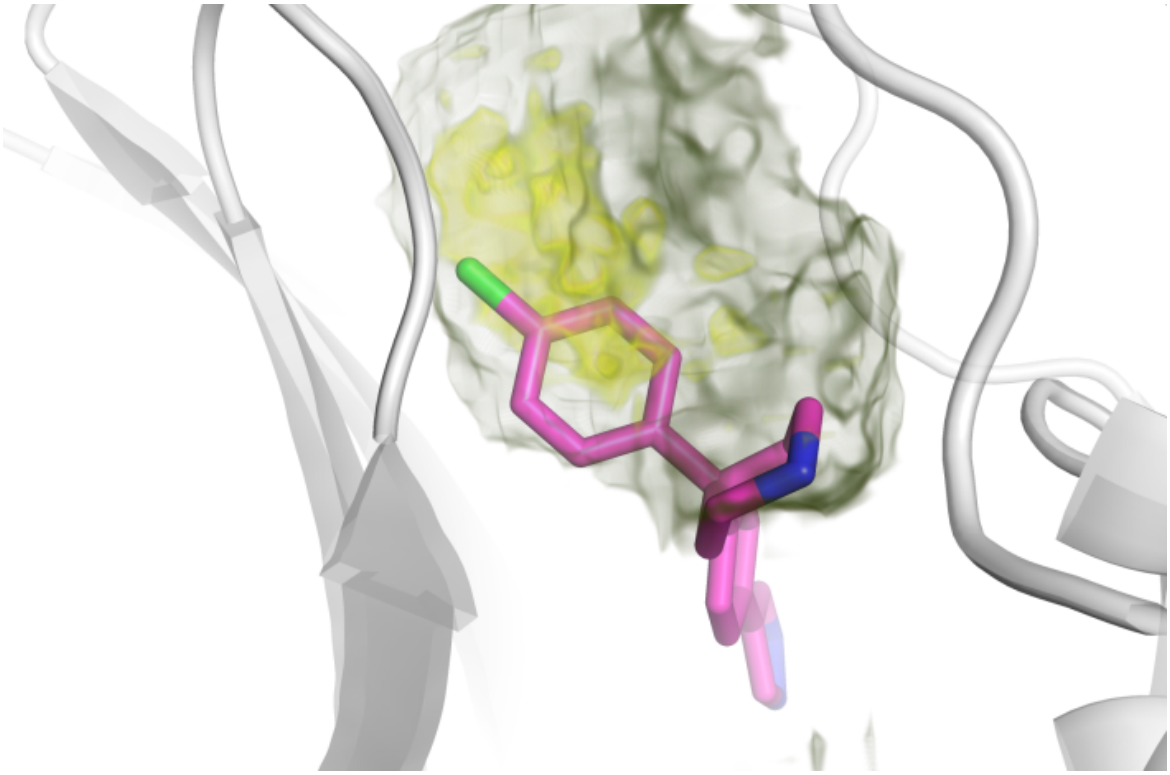


Fig. 3.7 Hydrophobic Fragment Hotspot Map calculated with the fragment left in the binding site, which was defined prior to the calculation. This allows for finer sampling of the pocket in much less time as a global search of the protein is no longer needed.

3.2.2 Pantothenate Synthetase

The lead molecule in this data set for pantothenate synthetase (PDB code: 3IUE) has been recently revisited [167], and the group efficiency (GE) of each group determined to highlight which part of the molecule should be developed further. Figure 3.8 shows how the GE is distributed throughout the molecule and how each intermediate structure compares to the Fragment Hotspot Map.

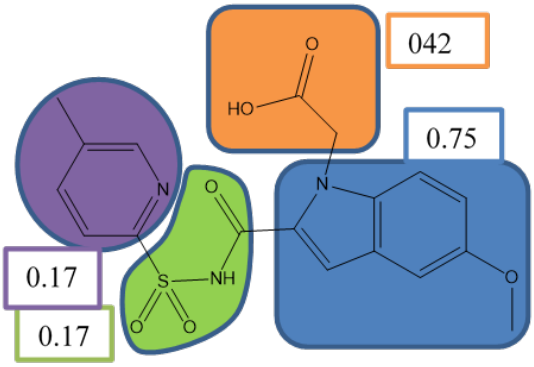
The fragment hit shown in figure 3.8b makes two specific interactions, the methoxy oxygen matches an area of high acceptor propensity and the NH forms a hydrogen bond with a sulfate ion. If the sulfate ion is included in the Fragment Hotspot Map calculation, then the NH is located in a region of high donor propensity. However, although present in the crystallization conditions, the sulfate is not present in the isothermal titration calorimetry (ITC) measurement and therefore does not help rationalize the high group efficiency of the fragment. It is possible that there is another ion in the ITC experiment that is able to bridge the interaction in the same manner as the sulfate. One possibility is 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), a buffer used in the ITC experiment. This is capable of making the interactions of the sulfate ion, in addition extending into the second hotspot (figure 3.9).

Addition of the sulfonamide moiety shown in figure 3.8c has a GE of just 0.17. This is perhaps surprising as one oxygen of the sulfonamide is placed in a very high scoring region and the NH forms a hydrogen bond with a bridging water molecule interacting with Gly-158. This water molecule is displaced upon binding of the transition state analogue shown in figure 3.11, so it is possible that this water is not particularly tightly bound, resulting in the lower GE. The remaining sulfoxide and carbonyl oxygen atoms are left interacting with the solvent, therefore again not contributing to the binding energy and reducing the GE.

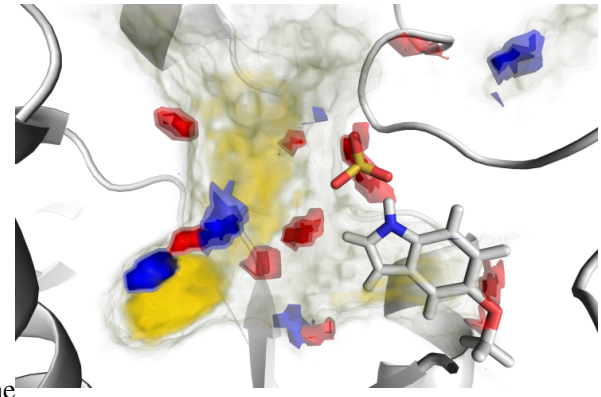
The methylpyridine moiety added in figure 3.8d is also not very group efficient. The ring occupies a moderately high scoring hydrophobic pocket but fails to satisfy any of the polar interactions of the protein or the acceptor nitrogen of the pyridine.

The final addition of the ethanoic acid group is highly group efficient. The oxygen atoms are placed in an area of very high acceptor propensity, resulting in a strong increase in binding energy for a small increase in atom count.

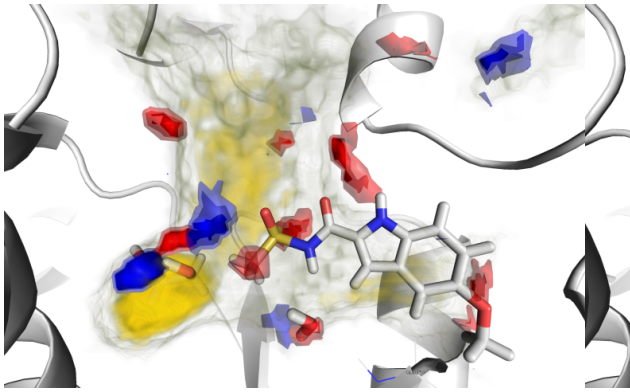
From the GE analysis, the methylpyridine group was highlighted as the best place for optimization. Toluene was found to be the most group efficient change, with a GE of 0.35 compared to 0.17. However, the crystal structures of the new more potent compounds showed



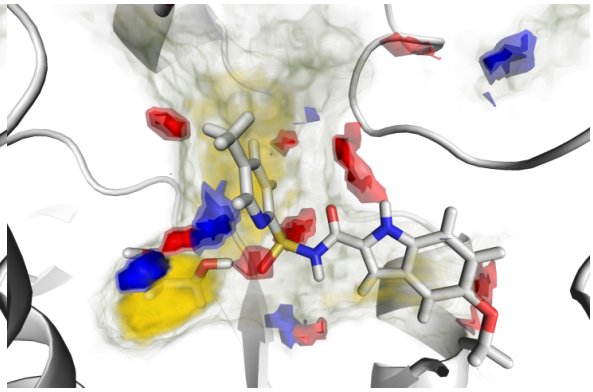
(a) Group efficiency for the groups that make up the lead molecule for pantothenate synthetase (PDB code 3IUE) [167].



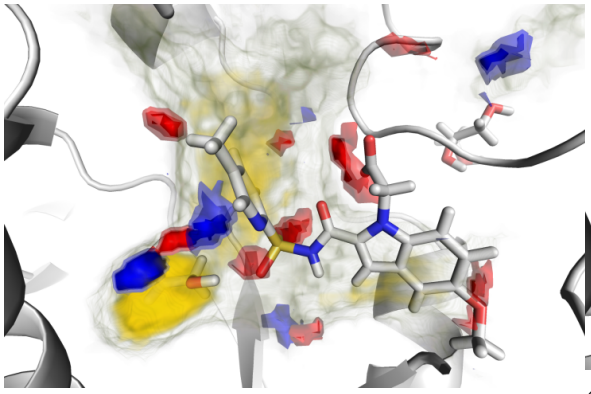
(b) Fragment hit (PDB 3IMG).



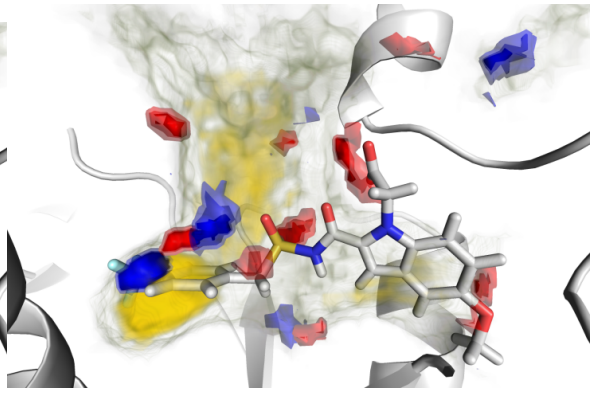
(c) Iteration of the hit-to-lead development used to calculate the GE (PDB 3ISJ).



(d) Iteration of the hit-to-lead development used to calculate the GE (PDB 3IUB).



(e) Iteration of the hit-to-lead development used to calculate the GE (PDB 3IUE).



(f) Recent inhibitor of pantothenate synthetase (PDB code: 4MUK), which fills the more highly scoring P1 pocket.

Fig. 3.8 Breakdown of pantothenate synthetase's GE. Hydrophobic map is shown in dark-grey to yellow to show moderate to high scoring regions. Donor hotspots are shown as a blue surface, and acceptor hotspots are shown as a red surface. The atoms seem displaced from the maps as the global alignment of the proteins did not manage to align the binding site well.

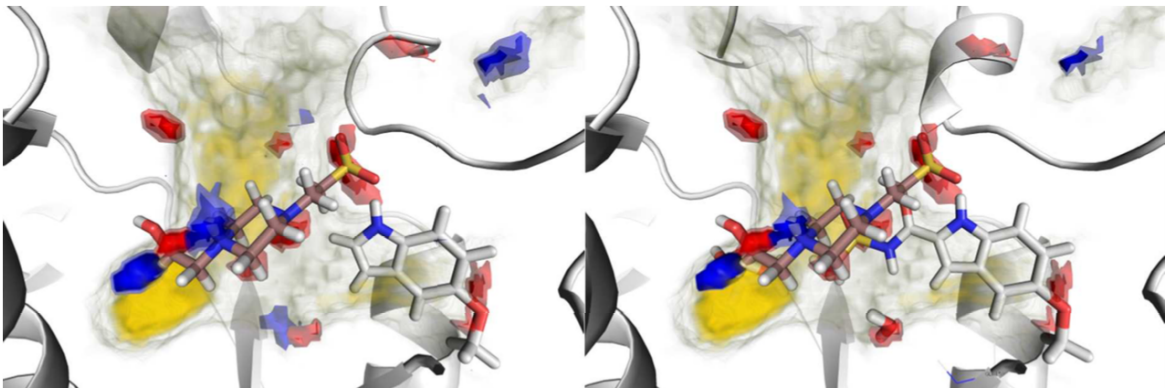


Fig. 3.9 Pantothenate synthetase with HEPES bound in place of the sulfate. This structure was created by aligning the sulfate of HEPES to the sulfate ion found in the crystal structure. This results in the HEPES molecule extending into the second hotspot.

that they no longer bound to the P2 pocket but instead bound to the P1 pocket, as shown in figure 3.8e. The molecule was developed further by adding a trifluoromethyl group and an extra carbon between the ring and the sulfonamide linker to completely fill the P1 pocket. The molecule has a reported IC_{50} of 250 nM. As is clear in the figure, the P1 pocket is a hydrophobic hotspot and is also capable of binding fragments.

The P1 pocket does have polar interactions available, but the lead molecule does not interact with any of these. Looking again at the percentage ranking of atoms across the whole data set, splitting the atoms into their corresponding interaction types shows the apolar atoms to rank more highly than donor and acceptor atoms for both fragments and leads (figure 3.10). In most cases, a few very highly scoring apolar regions correlated strongly with fragment binding locations. Polar interactions were more likely to be left unsatisfied by the fragments, but this does not mean they were unimportant for binding. The fragments in the data set were typically flat and unable to match all the interactions highlighted by the Fragment Hotspot Maps, exemplified in figure 3.11.

In contrast, when the maps were compared to a transition state analogue, almost all of the predicted high scoring interactions were satisfied. This is in line with the findings by Higueruelo *et al* [168], who used scissor plots [169] to describe the interactions of both synthetic and natural ligands in proteins. Natural ligands were found to maintain a greater ratio of polar to apolar contacts, with the number of polar atoms in the ligand correlating with the number of polar contacts. In contrast, synthetic ligands tended to find a few polar interactions with the remaining heteroatoms unmatched by the protein. For the lead atoms

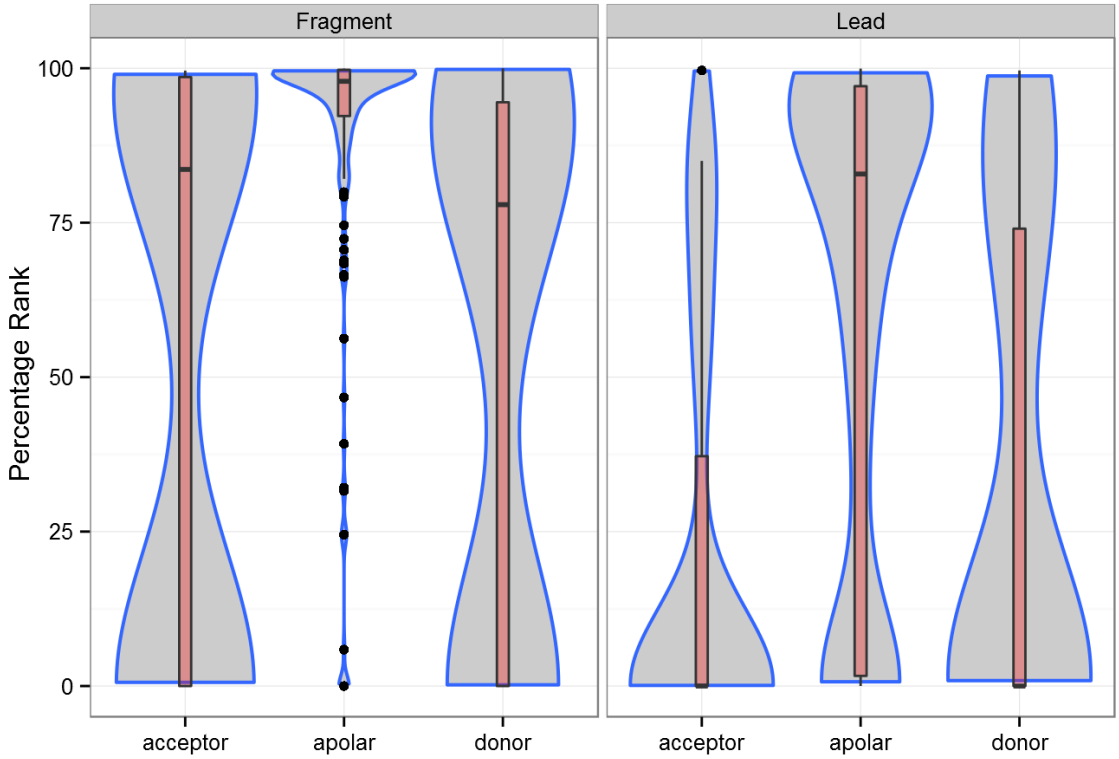


Fig. 3.10 Box and violin plots showing the percentage rank for fragment and lead atoms split by interaction type. For both fragment and lead atoms, the apolar atoms are the most highly ranking. Very few polar lead atoms were found to reside in highly scoring areas.

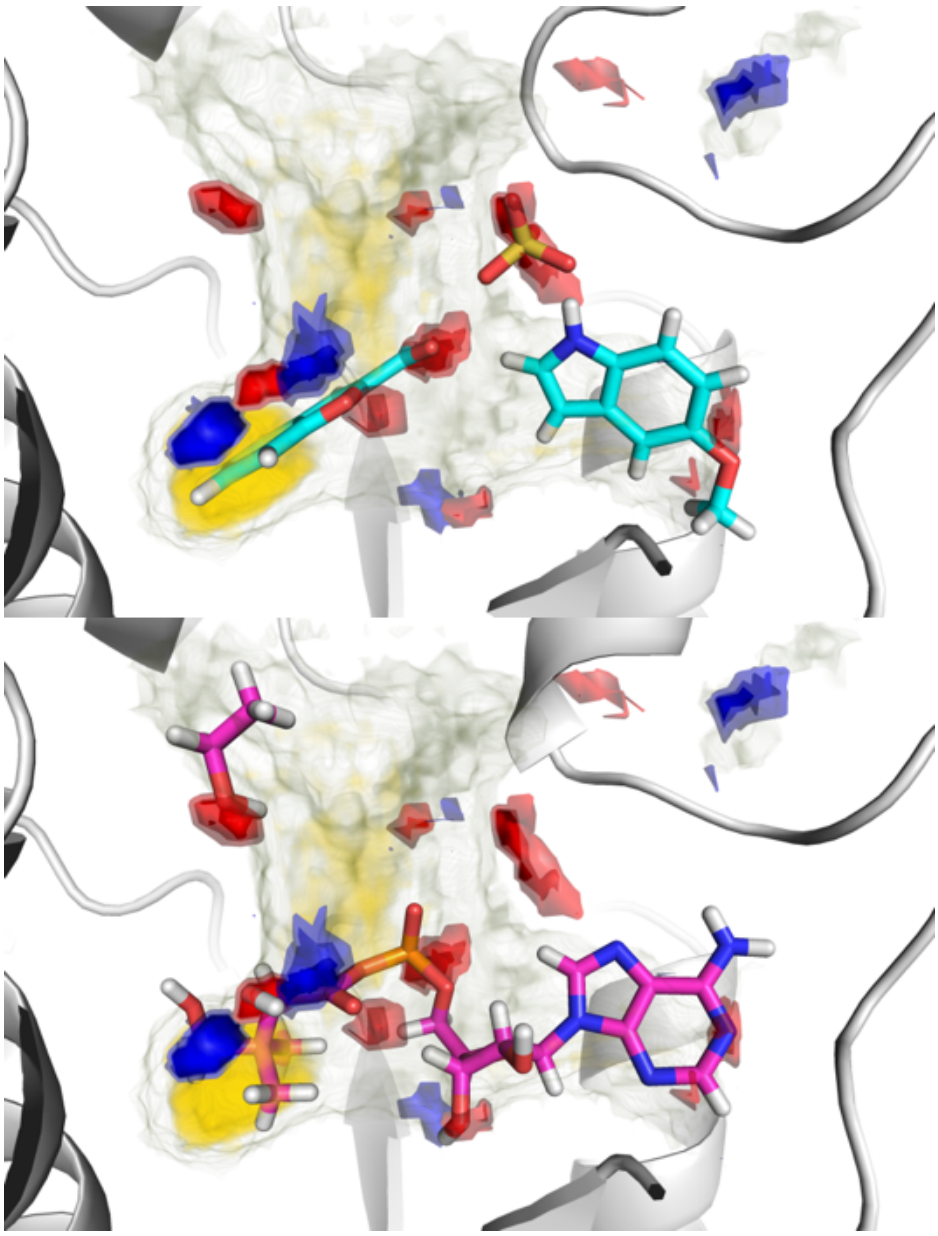


Fig. 3.11 Pantothenate synthetase showing (left) two bound fragments in cyan (PDB code: 3img) and (right) a transition state analogue in magenta (PDB code: 1n2h). Hydrophobic map is shown in dark-grey to yellow to show moderate to high scoring regions. Donor hotspots are shown as a blue surface, and acceptor hotspots are shown as a red surface. Although all the interactions of the fragments are satisfied, they leave many of the protein interactions unsatisfied. The more 3D and flexible transition state analogue satisfies the majority of interactions predicted by the Fragment Hotspot Maps. One interaction not satisfied in either case is found to bind an ethanol molecule (in magenta sticks), suggesting this is still a genuine hotspot interaction.

in our data set, relatively few polar atoms scored above zero. This could be a physical illustration of Hann's [66] complexity constraint or the result of the tendency in medicinal chemistry programs to add hydrophobicity to increase potency.

The two fragments bound to pantothenate synthetase each have two polar groups, both of which are satisfied. When using X-ray crystallography for a primary fragment screen, having a smaller library of simpler fragments like these is feasible as the elaboration of the fragments can be visualized in three dimensions; the structural data allows complexity to be built into the molecule to gain more affinity and selectivity by rational design during lead optimization.

However, some opt for larger collections of more complex fragments, for which X-ray crystallography is too low throughput. Pharmacophores derived from the Fragment Hotspot Maps could make the use of X-ray crystallography with these libraries feasible. Rather than relying on the promiscuity of simple fragments, the larger library could be filtered to give a subset that matches the hotspot interactions of the target. This may yield higher quality and more polar hits that would be more forgiving if hydrophobicity was then added to increase potency.

3.3 Conclusions

Identification of hotspots and their specific interactions can be used to evaluate the ligandability of a pocket and suggest which interactions fragments and larger ligands will need to make to bind there. Current methods use hotspots to assess either ligandability of subpockets from a global search of the protein [110], or provide interaction information for a predefined binding site [143].

The Fragment Hotspot Maps method is capable of providing interaction information from a global search of an *apo* protein crystal structure. As this method does not rely on MD, results can be calculated within minutes rather than hours on an ordinary laptop. Ligand atoms were consistently found in the highest scoring grid points; fragment atoms had a median percentage rank of 97% and lead atoms 72%. In addition to being able to identify fragment binding sites, the Fragment Hotspot Map method is able to highlight the interactions likely to be made by fragments. This makes the method useful at multiple stages in the drug discovery process.

The maps will complement existing virtual screening methods. With the most important interactions highlighted, existing pharmacophore methods can be used to screen for molecules capable of making these essential interactions. Equally, the maps can be used to generate constraints for docking, steering the docking toward occupying the hotspot and ensuring the right interactions are made. Initially, this required the maps to be visually inspected and then the docking constraints or structure-based pharmacophores to be generated manually, however, automatic work flows will be discussed in the next chapter.

Chapter 4

Improving Accessibility and Integrating with Existing SBDD Work Flows

4.1 Fragment Hotspot Maps Web App

4.1.1 Introduction

At the time of its publication [1], Fragment Hotspot Maps were calculated using a Python script that could be run from the command line. The script required an input protein that had been preprepared as described in chapter 2. It would generate a directory containing the output files along with a second Python script to help display the results (example shown in figure 4.1) in PyMOL [170], a Python-based molecular visualisation system. The PyMOL session would display the Fragment Hotspot Maps as isosurfaces contoured at 17, 14 and 10, values corresponding to "strong hotspot", "potential hotspot" and "binding site" as discussed in chapter 3. Little user input was required to calculate and visualise the Fragment Hotspot Maps, however the implementation limited who could run the calculations and how the results could be used.

Many people are unfamiliar with working from the command line, or may not have the required software (SuperStar and the CSD Python Application Programming Interface (API)) to be able to run the calculations on their own computer. One of the strengths of the Fragment Hotspot Maps method is the simplicity of the output, providing an easy-to-interpret visualisation of hotspots that does not require expertise in CADD. In a post on the Practical Fragments blog by Dan Erlanson [171] discussing Fragment Hotspot Maps, Erlanson concludes with the following:

Fig. 4.1 Example PyMOL session of results. Surfaces are pre contoured at 17, 14 and 10, with hydrophobe maps set to yellow, donor to blue and acceptor to red

> "Unfortunately, as currently described, the process will require a skilled modeller. It would be nice if the authors built a simple web-based interface for people to upload pdb files for analysis, as is the case for FTMap."

A web application has since been developed to meet these needs. With the aim of making the method as accessible as possible, the following requirements were set out:

- Inputs
  - Enter a PDB code, then retrieve the structure
  - Upload a PDB file
- Preparation
  - Allow user to start from an unprocessed PDB file
  - Give options to prepare the protein for calculation, such as adding hydrogens or removing waters
  - Default settings should reflect the most common usage
- Results
  - Visualise the results through the web app, or download PyMOL session for future use
  - Basic functionality for changing protein representation, without cluttering the interface
  - Intuitive method for changing the isosurface contour level, which gives a clear indication of what the score represents

4.1.2 Tools

Web development can be separated into two main areas: front-end and back-end. Front-end refers to the creation of the web page that the user sees and interacts with. It is important that the page is presented in such a way that the workflow is intuitive and consistent with what the user is accustomed to. Luckily, frameworks such as Bootstrap [172] are becoming

increasingly ubiquitous in front-end development. They help developers create professional-looking and familiar interfaces, making navigation of a new website feel instinctive. Back-end development refers to the logic, database interaction and calculations that occur in the background. In the context of the Fragment Hotspot Maps web app, this will include the retrieval and processing of PDB files, calculation of Fragment Hotspot Maps and storage of the results. There are a variety of back-end programming languages available, including python, which allows for easy integration with the Fragment Hotspot Maps scripts.

Pre-existing tools and frameworks allow scientists with limited web development experience to create intuitive websites within reasonable time frames. Further to this, freely-available protein visualisation tools are available to be embedded into web pages, allowing for results to be viewed directly on the website. Below is a list of the tools used in the development of the fragment hotspots web app:

Pyramid — A python-based web framework to run the back-end of the web-application. This controls how the server responds to the user's interaction with the site.

SQLalchemy — A Python Structured Query Language (SQL) toolkit for interaction with the database.

Jinja2 — An HyperText Mark-up Language (HTML) templating language for use with Python.

Bootstrap — Open source HTML, Cascading Style Sheets (CSS) and JavaScript (JS) framework for designing responsive and mobile-friendly web pages

jQuery — A JS library to help with HTML document traversal, event handling and animation.

NGL viewer — [173] A WebGL-based web-application for molecular visualisation

4.1.3 Work Flow

This section will describe the process of running a Fragment Hotspot Maps calculation. Each step will be described in terms of both the user experience, and work performed in the background.

4.1.3.1 Protein Input

User Experience Upon arrival to http://fragment-hotspot-maps.ccdc.cam.ac.uk/, the user is greeted with the page shown in figure 4.2a. There are two options for uploading a PDB file: Entering a valid PDB code, or uploading a valid PDB file. Optionally, the user can include a name to find their results later. If a valid PDB code or file has been provided, clicking submit will progress to the protein preparation page, else an error box will appear stating the PDB code or file is invalid.

Background Work When a PDB code is used as input, the PDB's File Transfer Protocol (FTP) is used to retrieve the corresponding file. If a file does not exist, the input page is reloaded with an additional message stating that the previous code was not found. Similarly, when a PDB file is submitted the CSD Python API is used to parse the PDB file, if it is unable to do so, an error is shown stating that the file is invalid. If the file or PDB code is valid, it is saved to the hard disk in a newly created directory, and the CSD Python API is used to ascertain which chains are present. A database entry is created for the job, and updated with the location of the saved PDB file.

4.1.3.2 Protein Preparation

User Experience The user is presented with a few basic options, and the asymmetric unit of their PDB file displayed in the NGL viewer, as shown in figure 4.2b. The first option is a drop down menu with each of the chain identifiers. It is possible to select any number of chains, and the protein's cartoon representation has been coloured by chain identifier to aid this selection. Tooltips are indicated by a "⃝?", and hovering the mouse over them provides helpful guidance without cluttering the page. For the chain select, the tooltip explains that the user should not run the calculation on the whole asymmetric unit, but either the biological unit or individual chains. After the chain select, there are two tick boxes to give options of removing water molecules and adding hydrogens. These are set to "true" by default, but can be disabled if the user has used alternative software to add hydrogens or wants to include some water molecules in the calculation. Tooltips are used to explain when this is appropriate. As the website is provided as a testing service, its computational power is limited and only two jobs can be run simultaneously. Very large proteins can lead to large memory usage and much longer calculation times, therefore a limit of 1000 residues has been put in place. If all the selected chains exceed this limit, an error is given and the user returned to the upload

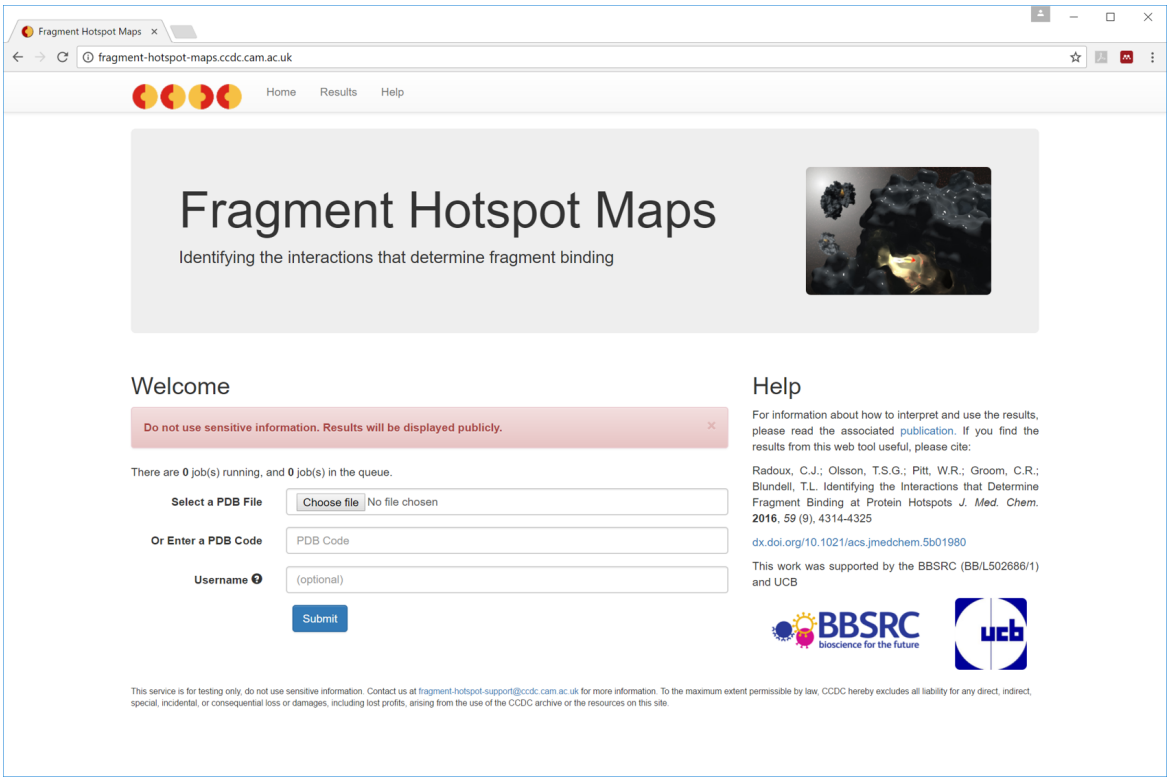
page. Otherwise, submission of the options results in only a small delay as the protein size is determined, followed by a redirect to the results table.

Background Work Once a PDB file has been loaded as a CSD python API protein object, it is reasonably simple to manipulate the protein to match the user's protein preparation settings. Once the prepared protein has been saved, the fragment hotspot calculation is ready to begin. There are two important considerations at this point. Firstly, the user's browser needs to still be responsive once the calculation has started. In order to achieve this, the calculation needs to be done asynchronously, meaning a new process is spawned for the calculation and the current process is allowed to redirect the user instantly to the next page. Secondly, a queuing system must be implemented to allow several jobs to be submitted and processed in order of submission, rather than all jobs running at the same time. The status of each job is tracked in the database, and can be set to "Queuing", "Running", "Failed" or "Complete". On submission of a new job, the number of running jobs is checked, if two are currently running, the new job is placed in the queue.

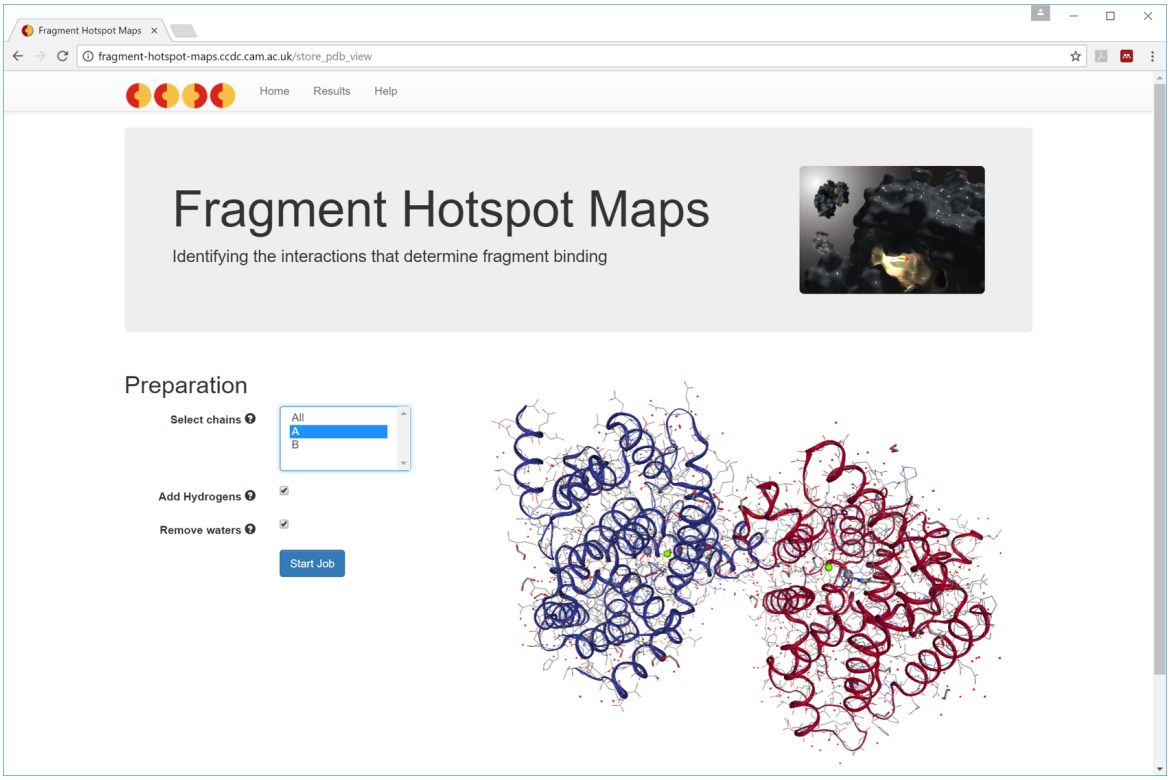
4.1.3.3 Results Table

User Experience Upon submission of a job, the user will be directed to the results table shown in figure 4.2c. The most recent job will be the first row of the table, and the status will be set to either "Running" or "Queuing". The table refreshes automatically, and once a job is completed the "view results" link becomes available and the next job in the queue is started. The table has a search box, allowing the user to search for a previous result by filename or user name.

Background Work The table is automatically updated through an Ajax (Asynchronous Javascript and XML) request. This allows the data in the table to be refreshed every 5 seconds, without having to reload the entire page. This process is only responsible for retrieving the table data from the database, and does not start the next job itself as this would mean new jobs only start when someone is viewing the results table. Instead, a server side process is constantly running and checking the number of running jobs. Once the number of concurrent running jobs goes below the maximum, the next job in the queue is started.

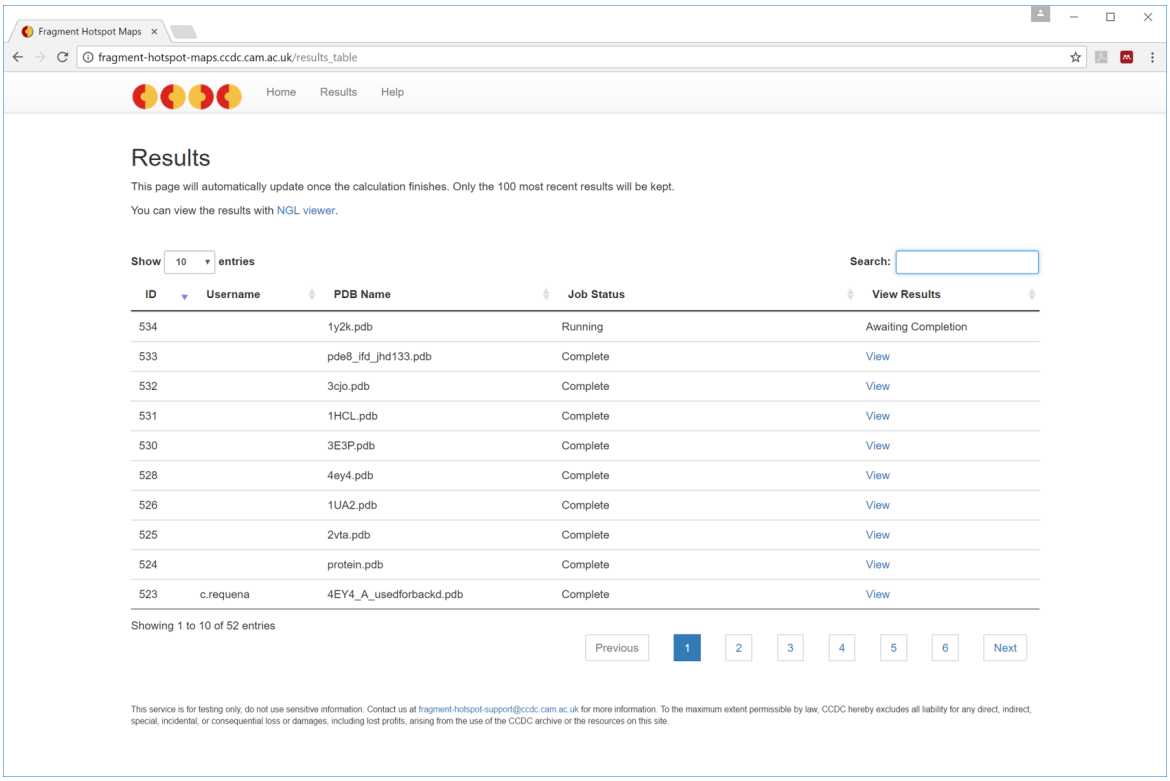


(a) Start page. Upload a PDB file or enter a valid PDB code.

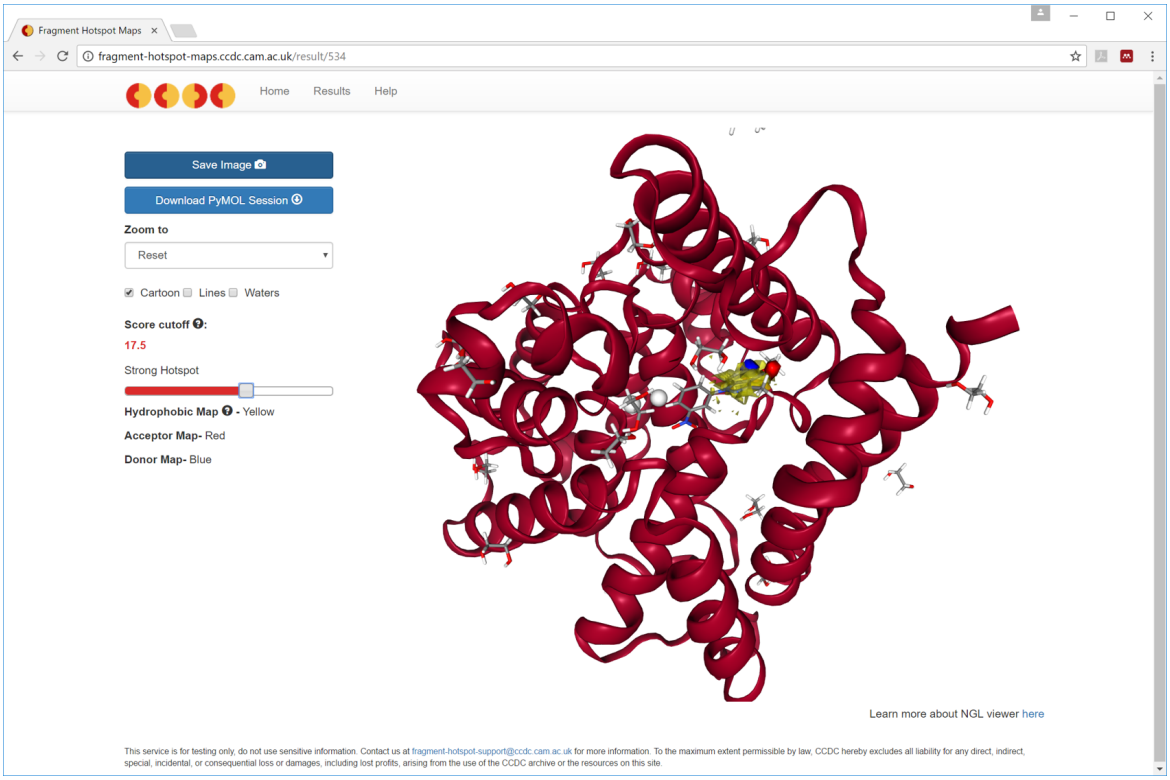


(b) Protein preparation page. Remove unnecessary chains, waters and add hydrogens if required.

Fig. 4.2 Using the Fragment Hotspot Web App



(c) Results table page. All results are displayed here, once the calculation is complete, the "view result" link will appear.



(d) Result page. Adjust the score contour with the slider, the maps change instantaneously with the movement.

Fig. 4.2 Using the Fragment Hotspot Web App

4.1.3.4 Result Page

User Experience The result page is shown in figure 4.2d. The left hand side of the window contains a selection of buttons, check boxes, sliders and drop down menus to aid visualisation and download results. The first two buttons allow you to create a high quality snapshot of your current view, or download a PyMOL session file containing the results. There is a drop down menu that allows the user to zoom and centre the view on one of the ligands found within the crystal structure, and three check boxes to toggle lines, cartoon and waters from the visualiser. The slider is used to control the contour level of the Fragment Hotspot Maps. Movement of this slider updates the contouring level instantaneously, updating the maps and the displayed score. To help the user understand what the score value represents, as the slider is moved a description of the current score is updated. Based on the values discussed in chapter 3, the scores are categorised as:

<10 Exposed interactions

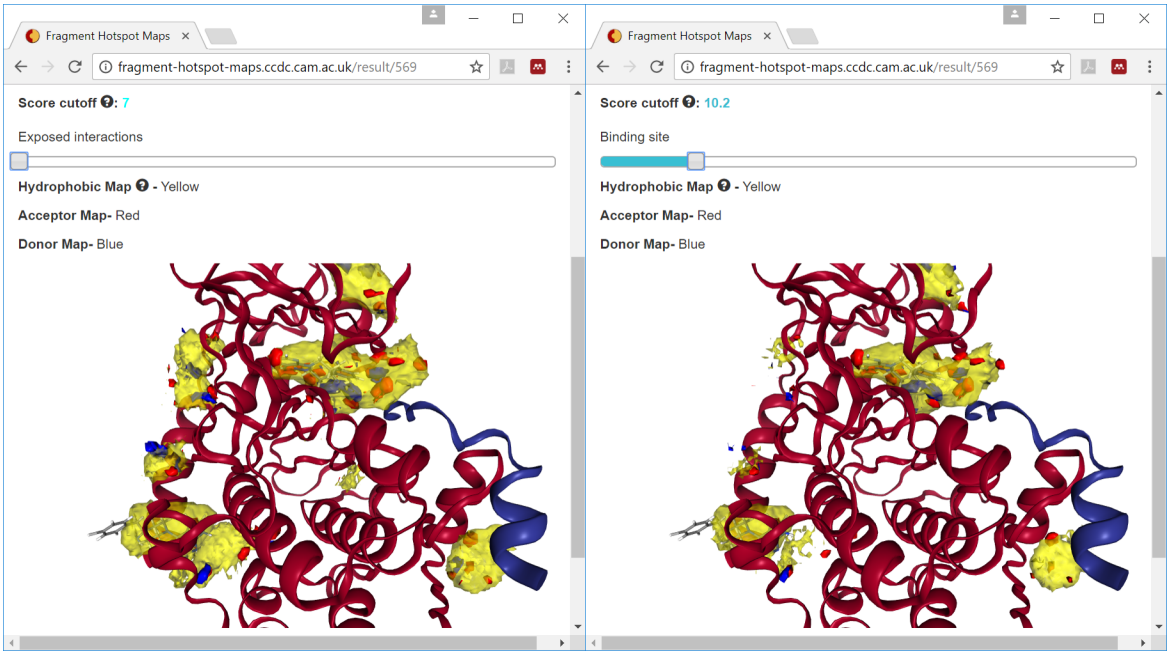
10-14 Binding Site

14-17 Hotspot

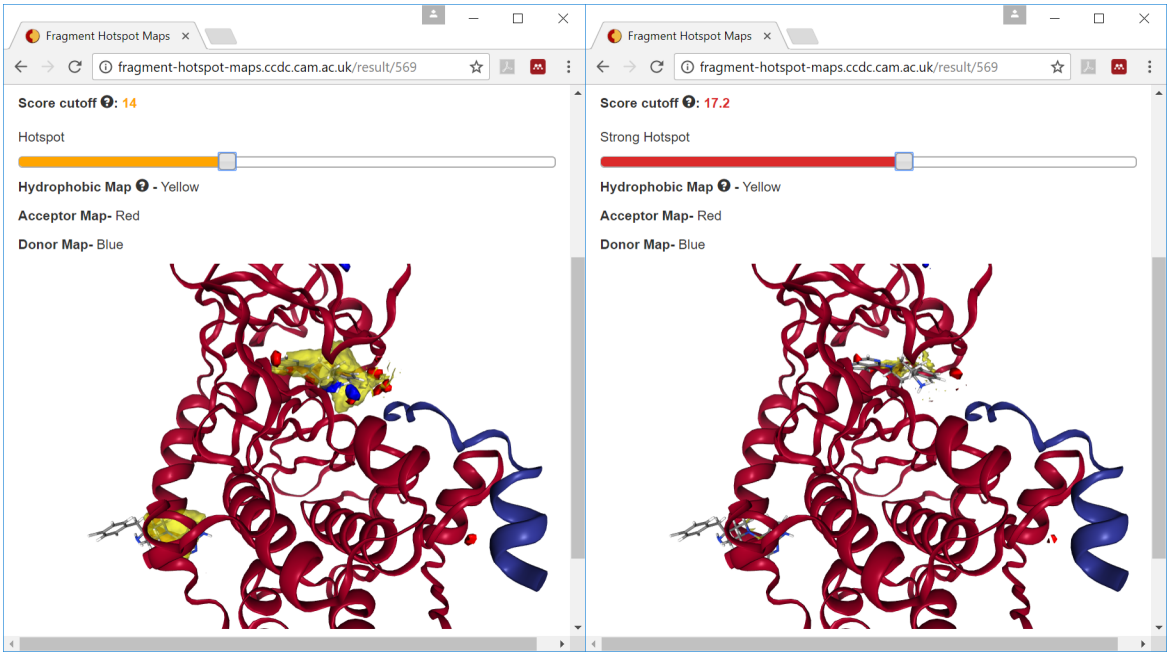
17+ Strong hotspots

The right hand side shows the protein and Fragment Hotspot Maps displayed with NGL viewer [173]. The user moves the slider to increasing score values until the Fragment Hotspot Maps cover a "fragment-sized" volume in the binding site as shown in figure 4.3. At this point, the value and the description from the slider can be used to assess the strength of the hotspot. The viewer is "responsive", meaning that its dimensions will match those of the window in which it occupies. This, in addition to the use of the Bootstrap framework, allows the results to be viewed comfortably on mobile devices and tablets. This is demonstrated in figure 4.3, where narrowing the window has caused the page to move to a vertical layout. The page contains tooltips for further explanation of how to interpret the maps, without cluttering the screen.

Background Work The interactive nature of the results page requires JS and jQuery. JS is an event-driven language, meaning that code is executed in response to the user interacting with the page. The NGL viewer is controlled through a JS API, allowing an event such as the movement of the slider to update the contouring level of the Fragment Hotspot Maps.



(a) Exposed interactions **(b)** Binding site



(c) Hotspot **(d)** Strong Hotspot

Fig. 4.3 Changing score contour with slider

4.1.4 Conclusion

The purpose of developing this web application was to remove any barriers for running Fragment Hotspot Map calculations. Providing the user has an input PDB code or file, calculations can be run with two mouse clicks, and results can be inspected using a simple slider to adjust the contouring level. As the contouring is adjusted, an assessment of the hotspot is updated and presented to the user, making outcome of the calculation clear. As a test of its simplicity, the web app was used by a group of second year undergraduates as part of a workshop titled "Getting to know your target". The aim of this workshop was to assess the tractability of a target, exploring data found in ChEMBL and the PDB as well as using online computational tools. Given the example of CDK2, all students were able to run the calculation and identify the hinge region as the ligandable site.

4.2 Hotspots API: Integration with the CSD Python API

Programming in science is becoming increasingly dominated by Python, a readable, general purpose language that runs on all major operating systems. This has led to the development of many specialist Python modules that allow users to perform complex tasks without having to write much code themselves. The CSD Python API is an example of such a tool, providing functionality to help with general handling of molecules and proteins, through to more CSD specific uses:

- Molecular file input and output API
- CSD Entry API
- Molecule API
- Searching API
- Conformer API
- Protein API
- Docking API
- Screening API

- Interaction API
- Descriptors API
- Diagram API
- Cavity API

The CSD Python API acts as a toolkit, Python scripts can be written to create work flows tailored to the specific task at hand. This is particularly useful for dealing with large datasets, where usage of a Graphical User Interface (GUI) would be repetitive and time-consuming for the user. This section will cover how the Fragment Hotspot Maps methods were rewritten as an API, and given functionality to create work flows that integrate well with the existing features of the CSD Python API. This work was made possible by my colleague at the CCDC, Richard Sykes. Richard is a Python developer, who has created much of the CSD Python API, and created additional functionality to handle grid objects such as the Fragment Hotspot Maps. This has greatly reduced the barrier for developing useful functionality for the Fragment Hotspot Maps.

There are two ways in which a "Hotspots API" will improve the efficiency of working with Fragment Hotspot Maps: How they are created, and how the results are utilised. The original command line implementation required a fully prepared PDB file as input. This required the use of other software, or running a separate CSD Python API script to first prepare the protein. The Hotspots API aims to allow the user to run a calculation in a number of ways from within a Python script, allowing both the preparation and calculation to be executed in one go.

A general overview of the hotspots API is shown in figure 4.4. The region highlighted in green shows the three methods available for running a fragment hotspot calculation. The central feature of this schematic is the "Hotspot_results object", shown in red. This is the result of the calculation, but requires further processing to generate a useful output. The parts highlighted in blue show how the results can be utilised, and any number of these functions can be used after Hotspot_results object is created.

4.2.1 Creating a Hotspot_results Object

There are three methods for generating a Hotspot_results object:

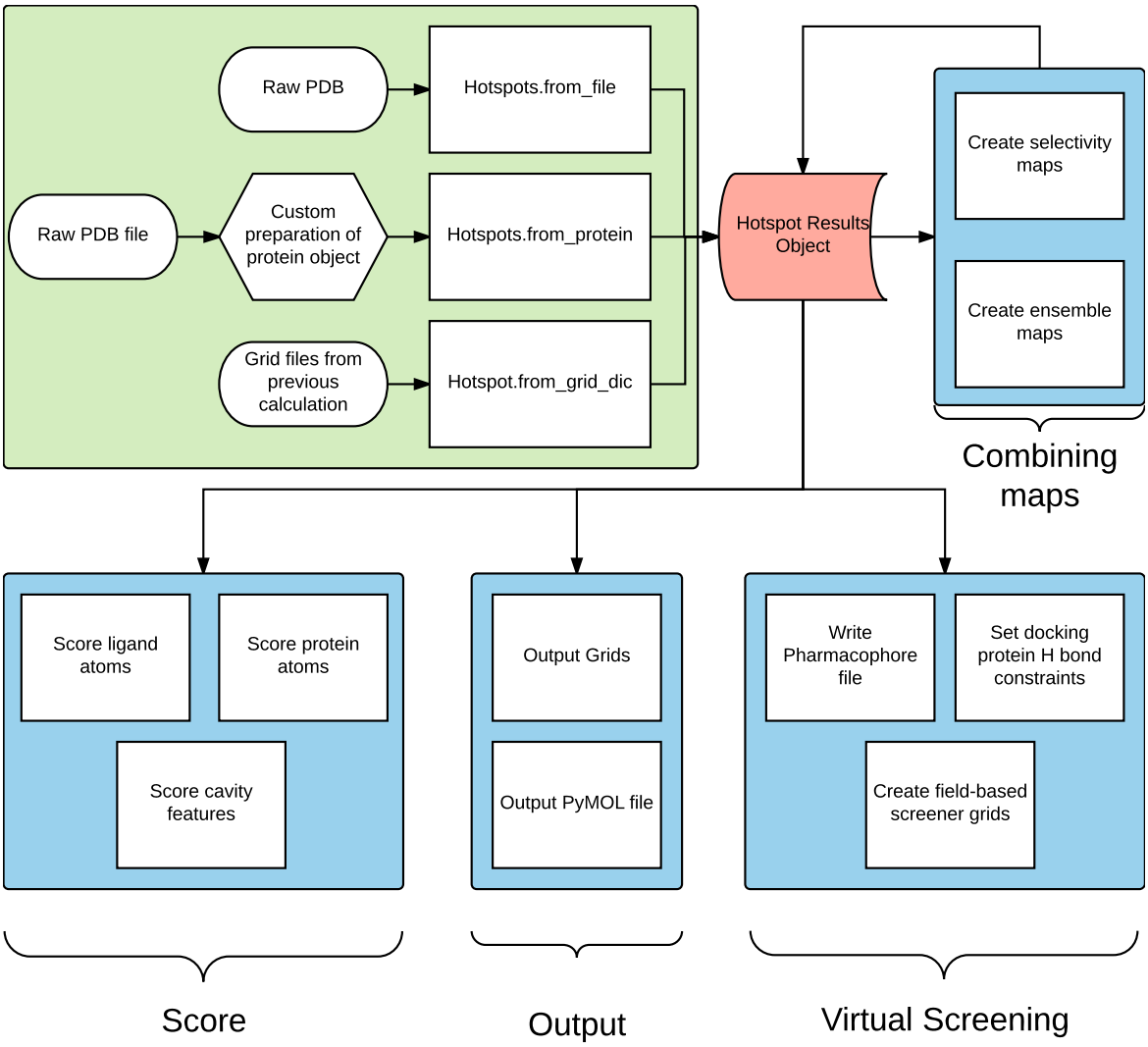


Fig. 4.4 Map of the Hotspots API. There are multiple methods for creating a Hotspot_results object (red), with different starting inputs (highlighted in green) . Once this object has been created, there are multiple functions available to utilise the results (highlighted in blue).

From protein object Calculate fragment hotspot maps from a CCDC protein object. The protein should have waters and ligands removed, and hydrogens added.

From PDB file Calculate fragment hotspot maps from a PDB file. Performs an initial faster cavity detection, running several smaller calculations and combining the outputs of each individual pocket. This is faster than running over the whole protein, but may result in some smaller pockets being missed. By default, this will prepare the protein by adding hydrogens, removing waters and ligands.

From grids Create a Hotspot_results object from a set of grid files. This allows all the functionality of the Hotspot_results object to be used with previously calculated results.

The three methods represent the different scenarios in which a user is likely to run a Fragment Hotspot Map calculation. Creating a Hotspot_results object from a file will be most useful at the start of a work flow, with optional arguments to prepare the protein. The user may also have previous results that they wish to revisit, and the "From grids" option allows grid files to be loaded directly into a hotspot results object to give them access to the functionality without waiting for the calculation to run once again.

Starting from a Protein() object allows the Fragment Hotspot Map calculation to take place within a greater workflow. An example is given in figure 4.5, which shows the overall process for a hotspot-guided docking run. This example starts with a GOLD docking configuration file (conf file) as input. The conf file contains the settings from a previous docking run, and loading it using the Docking API makes many of these setting accessible through python, including the Protein() object. This protein object can be used directly to create a Hotspot_results object, which is then used to highlight key hydrogen bonds in the binding site to guide the docking.

4.2.2 Using the Hotspot_results Object

As discussed in chapters one and two, fragment binding sites can give information about both the tractability of target and which interactions are likely to be essential within a binding site. As the Fragment Hotspot Maps method is capable of predicting where fragments will bind and which interactions they will make, this information can be used to help other CADD methods. This section will take a look at the different ways the Hotspot_results object can help streamline these work flows, concentrating on why each function will be useful and how it was implemented. A case study will be explored in detail.

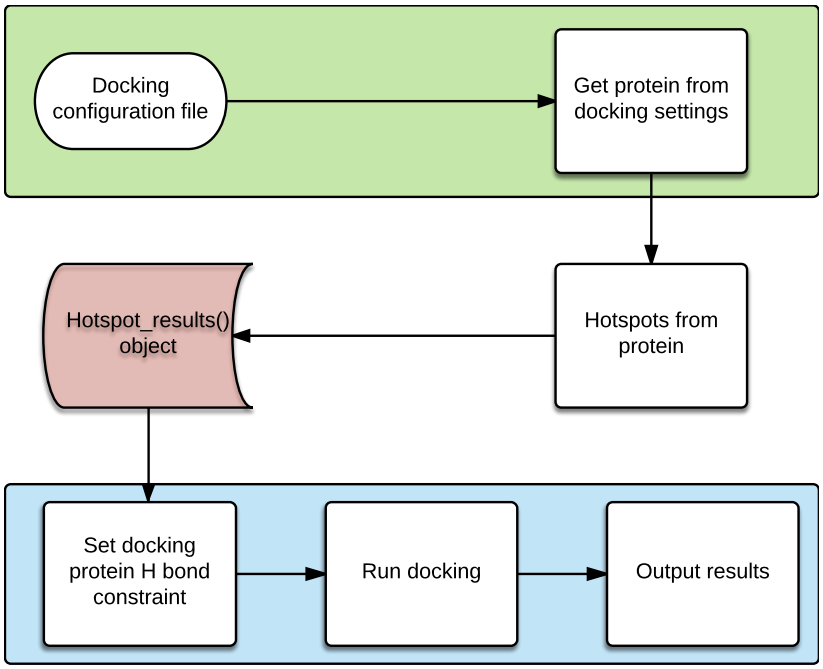


Fig. 4.5 An example workflow for hotspot-guided docking

4.2.2.1 Output

Get output grids This function returns the output grids produced by the hotspot calculation. These grids can then be saved to .acnt, .grd or .ccp4 files to be used later with other software or loaded in another script to crate a Hotspot_results object.

Output PyMOL File Outputs a Python script to be run from within PyMOL to visualise output. The function is flexible, and will allow the user to visualise anything that has been calculated by the user for the Hotspot_results object. For example, if the user has generated pharmacophores or results from a docking run, these can also be displayed in addition to the Fragment Hotspot Maps. The function will load all required input files, and set up the visualisation to be consistent with those generated in the publication.

4.2.2.2 Scoring

Each Fragment Hotspot Map is stored as a grid object, which contains functionality to lookup values at any point on the grid. The following functions describe various applications for reading the scores at a specific point in order to assign them to a given atom or feature.

Score protein – Assigns a score to each protein atom from its corresponding map (polar scores are taken from the partner atom type, i.e. a protein NH will get its score from the acceptor map). For polar interactions, an ideal interaction partner position is defined, and the highest scoring grid point within one grid point's distance is assigned to the protein atom. Donor scores are assigned to polar hydrogens, rather than the heavy atom. This is in line with the assignment of GOLD hydrogen-bond constraints, and allows scores to be assigned in cases such as a hydroxyl, which can act as a donor or acceptor.

Score ligand atoms – Assigns a score to each heavy atom based on type. For each atom, the highest score is chosen from the nearest grid point and all surround grid points, to allow for experimental error in the protein crystal structure. Returns a list of atomic scores. This is most useful in situations where the user would like to implement their own scoring method

Score Ligand – As above, but returns the geometric mean of atomic scores.

Score cavity features – The cavity API gives access to binding site comparison methods such as CavBase[174] and RAPMAD [175, 176]. The cavities are converted into a course-grained representation where the exposed residues are converted to a set of pseudocentres: hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor, hydrophobic aliphatic, and aromatic. This function assigns a score to each cavity feature from its corresponding map. This uses the same approach as scoring the protein atoms, but instead returns the cavity feature objects and their corresponding scores. This workflow will be discussed in detail in the case study below

4.2.2.3 Combining Maps

Common mathematical and logic operators can be applied grid objects in the CSD Python API. For example, $g1 * -1$ would multiply all points in grid g1 by -1 and $g1 - g2$ would subtract all points in g2 from g1. A prerequisite for mathematical operations involving multiple grids is that the grids have the same size, shape and origin.

Selectivity map – Using a second Hotspot_results object created for an off-target protein you wish to gain selectivity against, this function will return a new Hotspot_results object that corresponds to high scoring regions present in the target protein, but not in the off-target protein. The two proteins need to be aligned prior to the calculation, however this will still

result in grids of different size, shape and origin. This function first modifies the grids such that they have the same size, shape and origin, then subtracts the off-target grids from the target grids.

Ensemble map - Takes a list of Hotspot_results objects created from aligned protein structures and calculates the average maps across the ensemble. This will return a new Hotspot_results object that has access to all of the same functions. Much like the selectivity map, the grids from each protein are modified so that all maps across the set have the same size, shape and location. The mean is calculated by summing the maps then dividing by the number of proteins.

4.2.2.4 Virtual Screening

This final section describes some of the higher level functionality in the Hotspot API. These functions coordinate more complex work flows, whilst still only requiring a few lines of python from the user.

Write pharmacophore file – This function generates pharmacophore models from the highest scoring regions of the fragment hotspot maps. Islands of propensity above the given percentile (90th by default) are used to define pharmacophoric features. For polar maps, the centroid of the pharmacophoric feature is set to the highest scoring grid point within the island, for apolar maps, the feature is set to the centre of the island. Polar features within 6 Å of a given apolar feature are assigned to that pharmacophore. As a result, several pharmacophore files may be produced, corresponding to separate hotspots.

Predict protein H bond constraints – Takes a ccdc.docking.Docker.Settings instance and adds protein H bond constraints to the top n scoring polar interactions, where n is the number of constraints. A constraint instructs GOLD to penalise any docking pose that does not make an interaction with the selected protein atom. This function makes use of the "score protein" function to assign scores to each of the protein atom.

Ligand screener grids – Write out .acnt files that can be used by the ligand screener, which will be covered in more detail in the next chapter.

4.2.3 Case study: Hotspot-guided cavity comparison

An area of active research at the CCDC is storing and comparing protein cavities, using the surface shape and interaction features of a binding site. The ability to compare a pocket of interest to all other known pockets has fundamental applications in drug discovery:

- Identifying similar binding sites likely to give rise to toxicity or off-target effects
- De-convoluting polypharmacology
- Understanding the evolution of protein binding sites
- Elucidating the role of proteins of unknown function
- Mapping the plasticity of protein cavities in different apo and holo configurations
- Enhanced virtual screening ability
- Bioisosteric replacements of specific functional groups

Historically, sequence similarity has been applied to identify similar proteins; however, this is largely limited to similarity arising from a common evolutionary starting point. The comparison of binding cavity shape and features is much more powerful, but to date experiments have been limited by computationally intensive calculations. Queries may take from days to weeks to run, which restricts large scale calculations. New advances, however, can reduce this time to minutes without compromising quality [175, 176].

Future work will use the Fragment Hotspot method to prioritise features within a cavity, such that they are described by interactions that are highlighted as important for ligand binding. Combining the Hotspot, Cavity and Interaction APIs will allow two research areas to be explored.

4.2.3.1 Hotspot-guided cavity searching

Currently all cavity features, including those involved in hydrogen bonding with other protein residues, are considered equally during a cavity comparison. One area to be explored is the effect of reducing the features to include only high scoring interactions. A database of cavities will be created and decorated with hotspot regions as described in figure 4.6, and previous validation experiments will be repeated to assess the performance of cavity comparison by prioritising the hot areas of the binding sites. Once a list of potential off-targets have

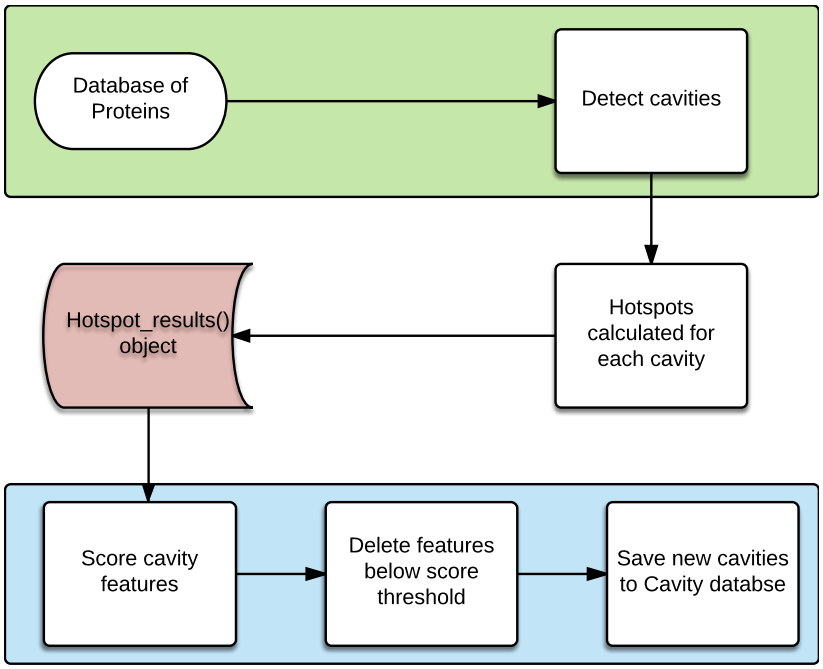


Fig. 4.6 Hotspot-guided cavity comparison

been identified, selectivity maps can then be created to suggest how to avoid binding to the off-target proteins.

4.2.3.2 Cavity vs ligand comparison algorithm

Changing focus to small molecules, FIMs[116] is a method based on SuperStar to generate maps that highlight the most frequent positions of interaction partners around a given molecule, exemplified in figure 4.7. As with Fragment Hotspot Maps, the FIMs use probes to show favourable donor, acceptor and aromatic probe positions. The peaks in the maps surrounding the small molecule represent ideal binding interactions for the particular conformation of the molecule, as predicted using interaction data in the CSD.

Reducing the cavity features to those likely to be involved in ligand binding may allow cavity comparison methods to take a ligand as a query. As summarised in figure 4.8, the peaks in the FIMs can be used to create a query binding site. This "binding site" can be compared to the dataset of hotspot-guided cavities using existing cavity comparison methods. This algorithm will be exploited to find putative binding sites from hits of phenotypic screens, and molecules displaying polypharmacology.

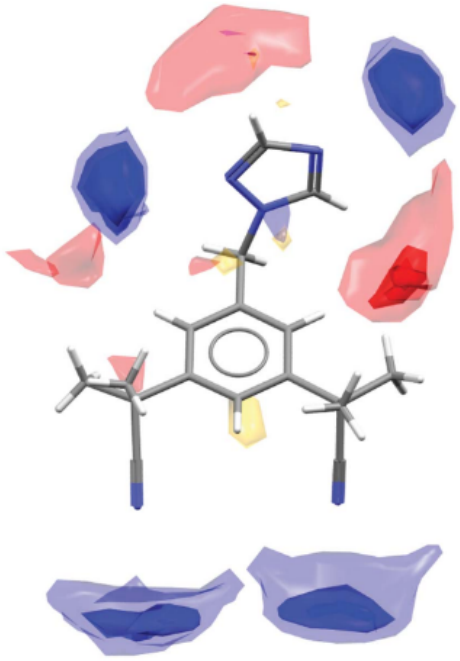


Fig. 4.7 Full Interaction Maps example, showing ideal positions for interaction partners based on data in the CSD. Hydrogen bond donor show in blue, acceptor in red and hydrophobe in yellow

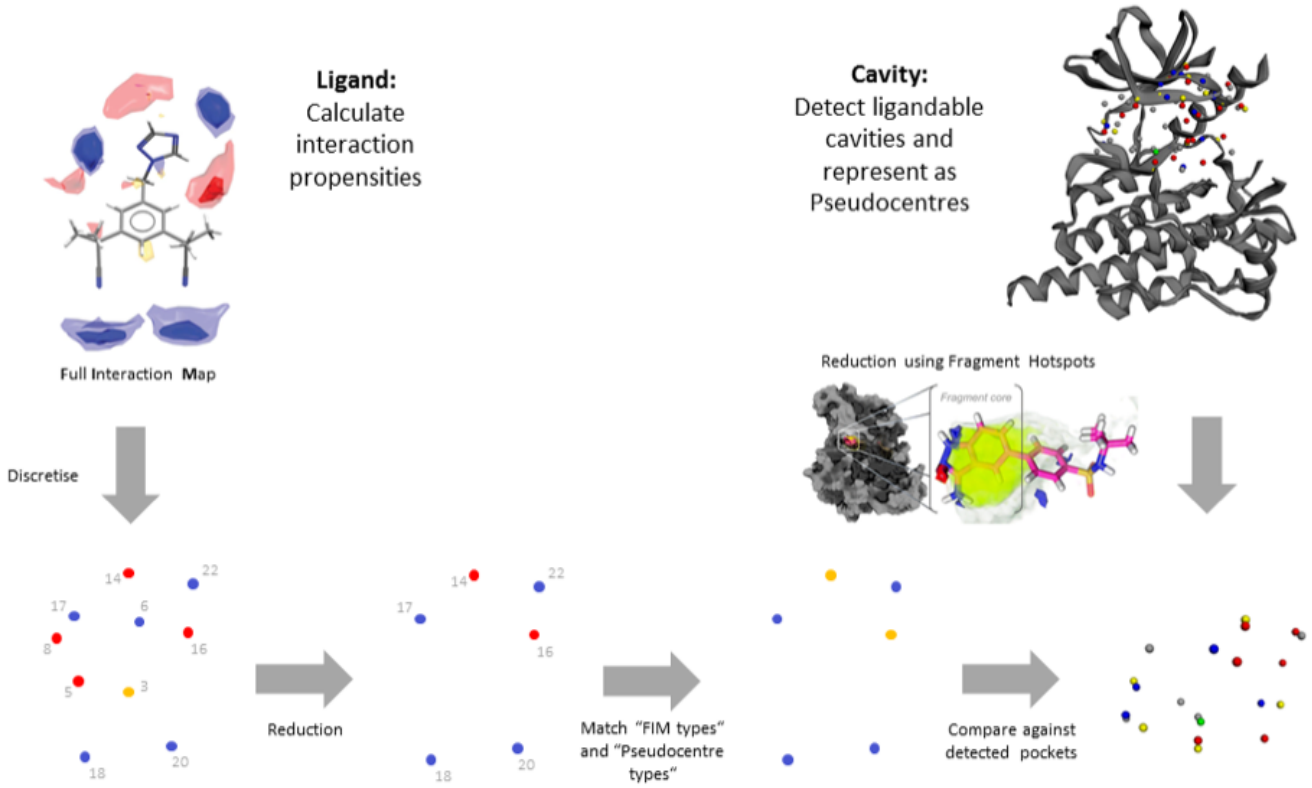


Fig. 4.8 Work flow for searching for targets of a given ligand. (Left) Full interaction maps are used to create a query binding site that match the ideal interactions predicted by FIMs. (Right) Reducing the detected cavity features to only include those with high scores from Fragment Hotspot Maps. Figure prepared by Timo Krotzky (unpublished)

4.2.4 Conclusion

The Hotspots API has provided a set of tools to grant greater flexibility in how Fragment Hotspot Maps are created and used. It gives researchers access to functionality that would otherwise be too time consuming to develop themselves as part of a wider project. This, in addition to the web application, has greatly improved the accessibility and utility of Fragment Hotspot Map calculations for people of all ranges of CADD and programming experience. Some features of the Hotspots API have not yet been validated scientifically, such as the hotspot-guided cavity comparison discussed in the case study, but have been developed in order to enable future research. Other areas, such as virtual screening, have been looked at in more detail, and will be discussed in the next chapter.

Chapter 5

Hotspot-Guided Virtual Screening

5.1 Introduction

For a small molecule to bind to a protein, there are several barriers that it must overcome. Some are relatively simple, such as the need for shape and interaction complementarity, however others are more complex. The ligand will face entropic penalties due to both its lost ability to rigidly tumble and rotate[99], and loss of conformational and vibrational degrees of freedom, which are more complex and typically require MD-based calculations[177]. Further to this, both the ligand and protein may face internal strain as they assume conformations suitable for the binding event. Any hydrogen bonds formed in the protein-ligand complex must first displace water molecules on both the ligand and protein, which can have drastically different effects on binding depending on their environment[132, 130]. Once the protein-ligand complex has formed, both the resulting network of waters surrounding the solvent-facing part of the ligand[178, 179] and the network of protein residues[180] have an impact on the binding free energy. The quality of a resulting protein-ligand complex should not be judged by the sum of its interactions[166], but rather by the overall free energy change of the system starting from a fully solvated protein and ligand, and ending with the binding complex. Although possible to predict computationally the absolute free energy of binding for drug-like molecules while considering the system as a whole, it is only possible at huge computational expense[181], 29 hours for each complex on 504 cores (Intel Xeon E5-2697 v2 2.7 GHz), and 7 hours on 372 cores for the ligand.

In early drug discovery, it is necessary to virtually screen hundreds of thousands of molecules ahead of a primary screen. As a result, approximations are required to to allow

for much more simple and computationally inexpensive methods that can handle such an input. These methods do not seek to accurately predict the binding free energy of a complex, but rather prioritise those compounds which are more likely to bind. There are multiple approaches (table 5.1), depending on the type of data available, however they can be broadly categorised as ligand-based or structure-based.

Ligand-based methods include both 2D and and 3D approaches. For 2D methods, molecules are represented as binary bit strings (fingerprints), where each bit represents a specific feature. Features present are set to 1, whilst those absent are set to 0. The similarity between two fingerprints can be calculated using a number of methods, the most common being the tanimoto coefficient [182] using equation 5.1 (C is the number of bits set in common in the query and the database structure, Q is the number of bits set to on in the query structure, and D is the number of bits set to on in the database structure). This can be used to query databases of molecules to find those similar to known actives.

Although they are good at identifying close analogues, one limitation of 2D methods is their inability to find novel chemistry capable of binding to the protein, exemplified in figure 5.1. This problem is circumvented in 3D ligand-based methods such as pharmacophoric screening, as the structures of actives molecules are turned into abstract pharmacophoric features (figure 5.2 created using Pharmit[183]). A match will be returned providing the database molecule is able to place matching atoms with each of the spheres, however the definitions of a match can be much looser and the scaffold connecting the functional groups can be completely different to the query molecule. A downside to ligand-based pharmacophoric screening is that it relies on having a conformation close to the bound conformation, requiring either a protein structure containing the active ligand or the use of a ligand overlay program[184] in conjunction with multiple actives.

It is also possible to generate pharmacophores using the protein structure, but choosing which interactions to use can prove difficult for novel pockets. As described in the previous chapter, it is possible to create pharmacophores from the Fragment Hotspot Maps themselves, shown in figure 5.3. By creating pharmacophoric features for just the interactions in the hotspot, the pharmacophore model can search for molecules that satisfy these essential interactions. Once a hit is found, the ligand can be scored in the Fragment Hotspot Map to identify how well the rest of the molecule matches the interactions. A big benefit of pharmacophore screening is speed, Pharmit is able to screen millions of molecules within seconds to minutes[183], depending on the complexity of the query. The use of hotspot-

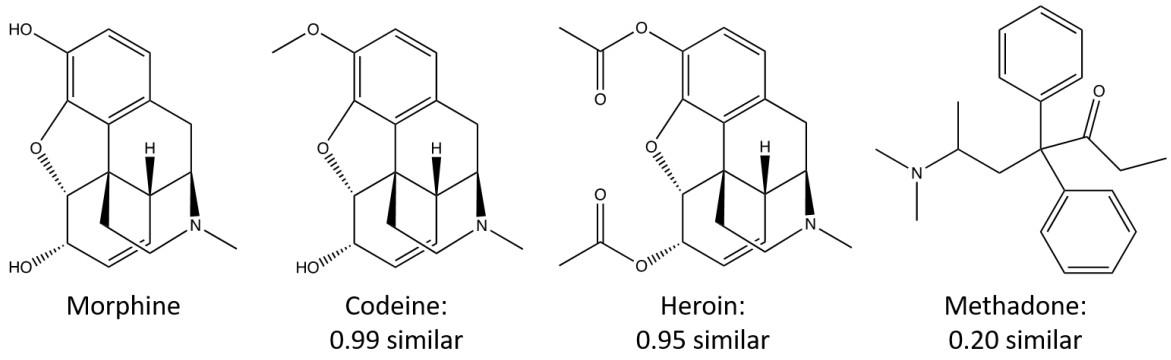


Fig. 5.1 Tanimoto coefficients calculated between Morphine and Codeine, Heroin and Methdone.

Table 5.1 Virtual screening methods

Method	Category	Example
Docking	Structure-based	GOLD[185]
Pharmacophore searching	Structure and/or ligand-based	Phase[186]
Shape-based screening	Ligand-based	ROCS[187]
2D Similarity searching	Ligand-based	Review of methods by Sheridan and Kearsley[188]

derived pharmacophores is ongoing research for another PhD student at the CCDC, Peter Curran, so will not be discussed in this chapter.

$$Similarity_{QD} = \frac{C}{Q+D-C} \tag{5.1}$$

The more complex barriers to binding discussed at the start of this chapter are typically too difficult to include explicitly in virtual screening. One exception is WScore in Glide docking[189], which uses the results from a WaterMap calculation to give a flexible description of explicit water molecules, leading to improved docking performance. While WScore uses MD-based WaterMap calculations to calculate the thermodynamic properties of hydration sites, a more simple approach will be taken here. Fragments are able to overcome all of these barriers despite having a limited binding interface with the protein, making highly efficient interactions at the hotspot. As these interactions also make a disproportionately large contribution to the free energy of binding in larger molecules, it is important that virtual screening methods favour making these interactions over others. In the absence of

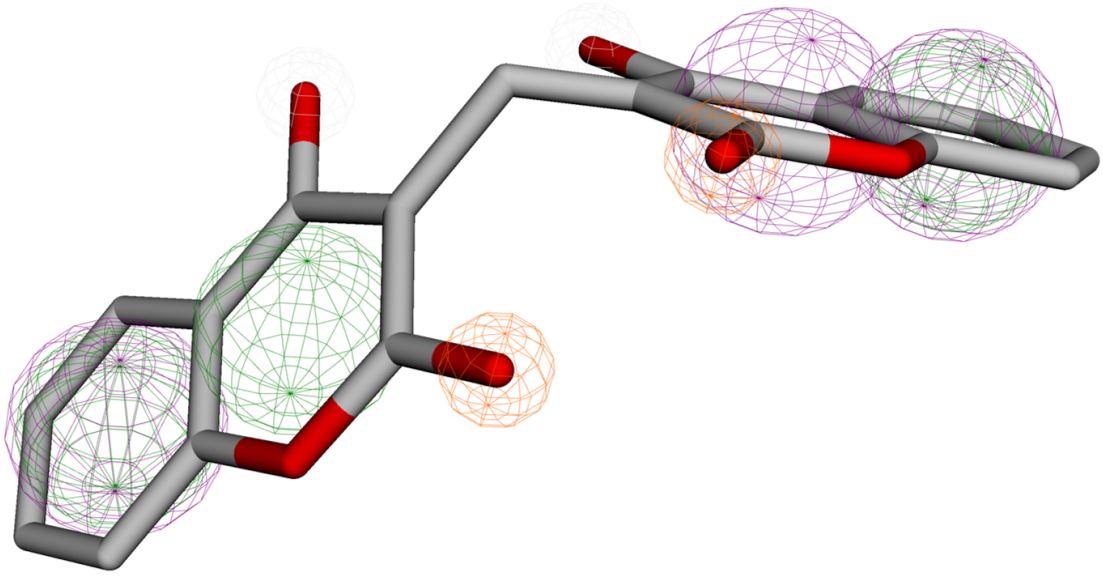


Fig. 5.2 A ligand and its corresponding pharmacophores

fragment-bound crystal structures, interactions at predicted fragment binding sites can be used.

This chapter will look at two virtual screening methods, and how information from Fragment Hotspot Map calculations can have an impact. The first approach will use docking with GOLD[185], and use the "predict protein H-bond constraints" method discussed in the previous chapter. The second approach looks at using a field-based virtual screening tool, which was designed as a 3D ligand-based tool, but has been modified to take Fragment Hotspot Maps as a direct input.

5.1.1 Docking with GOLD

Docking programs consist of two main parts: a method for sampling poses and a method for scoring these poses. Generation of ligand poses involves sampling its rotational and translational degrees of freedom as well as its conformational degrees of freedom. Brute force sampling is far too slow and would spend a lot of time sampling high energy structures, therefore methods have been developed to have reduced but more relevant sampling. While one of the earliest docking programs, DOCK[190], used rigid ligands to avoid this, GOLD[185] uses a genetic algorithm (GA) to allow full ligand flexibility.

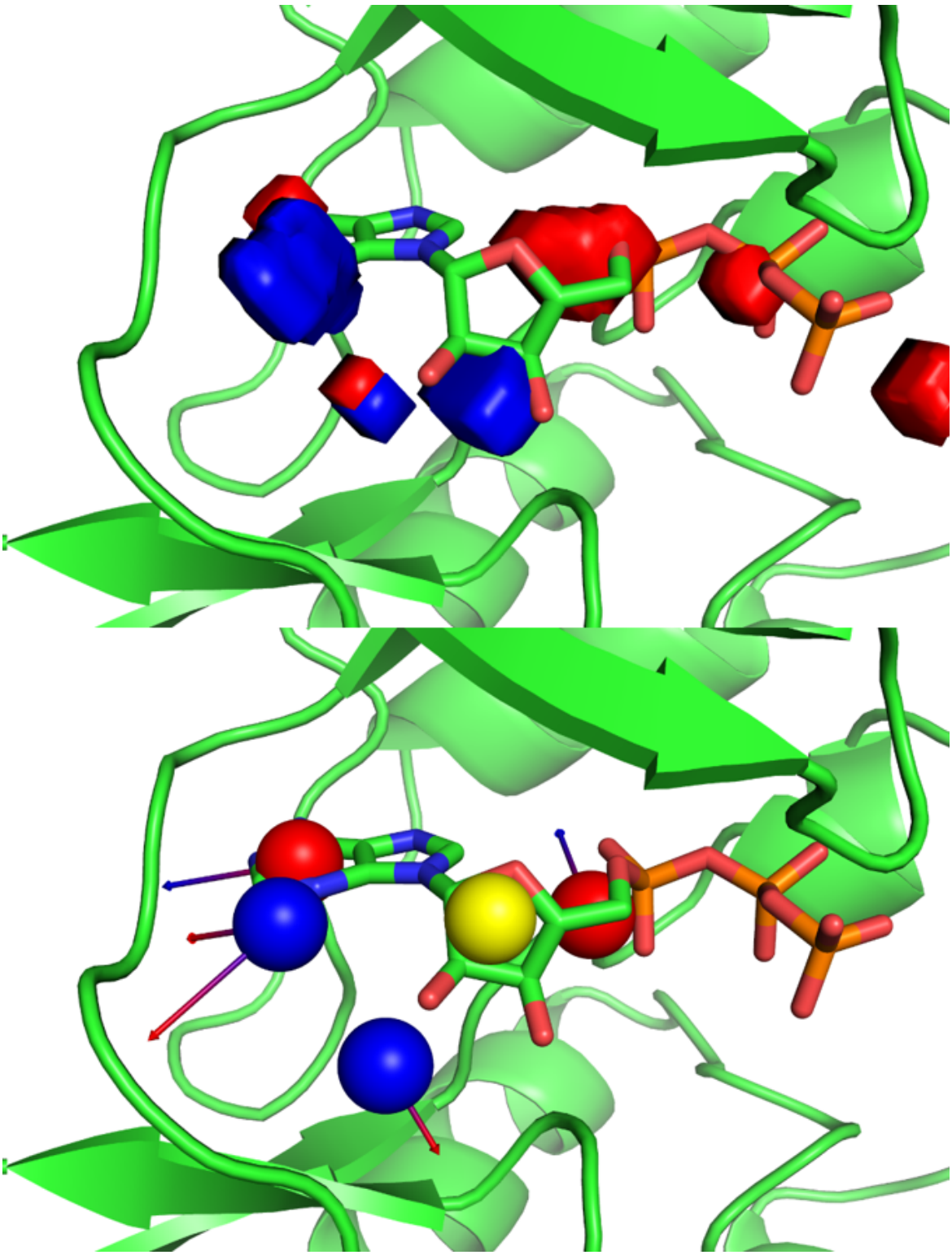


Fig. 5.3 CDK2 with pharmacophores generated from Fragment Hotspot maps and ATP displayed as sticks. Only the polar maps have been displayed for clarity.

As the name suggests, a GA requires a starting population that "evolves" to give the output poses. A ligand's rotatable dihedral angles and protein-ligand hydrogen bonds are represented as genes, which can undergo genetic operations. The collection of genes for a given pose is called a chromosome, and each chromosome is assigned a fitness score. Two parent chromosomes are selected, with a bias towards to those with greater fitness scores. Genetic operators are applied (crossovers, mutations and migrations) to the parent chromosomes and a new structure is generated. The number of inter-molecular hydrogen bonds are maximised using a least squares fitting protocol. The final structures are then ranked with a more complex scoring function. Since the original GOLD publication[185], hydrophobic fitting points have also been included as part of the least squares fitting protocol[191].

When performing a docking calculation on a pocket with known actives, it is possible to steer the docking towards the correct answer by using docking constraints. A protein hydrogen bond constraint will place a penalty to the fitness of a pose that does not make a hydrogen bond to the selected protein atom. This will not only favour the ranking of molecules that make the chosen hydrogen bond, but also guide the GA sampling to guide solutions towards those where the chosen interaction is made.

The importance of docking constraints is exemplified by the docking of S-adenosyl-L-homocysteine (SAH) into MLL1. SAH (figure 5.4) is a very flexible molecule, with 11 non-terminal rotatable bonds, and the pocket of MLL1 is very open, making this a difficult case. Under the default setting, GOLD is unable to correctly place the amino and carboxylate groups of SAH (figure 5.5a). Although the core of the molecule is correct, there is insufficient sampling for the flexible "tail" of the molecule to find the interaction. One solution is to greatly increase the sampling performed by GOLD, which is able to bind the correct pose, however the calculation time increases from approximately 3 minutes to over 15 minutes. Calculation of Fragment Hotpot Maps for MLL1 produces the output shown in figure 5.5d when contoured at 17. These are the only three interactions highlighted across the whole protein, with the region of blue donor propensity giving the highest score. Adding the histidine nitrogen as a constraint yields the correct pose within 3 minutes, as shown in figure 5.5c. Identification of the correct binding pose is important in virtual screening as it is required to correctly assess the ligand.

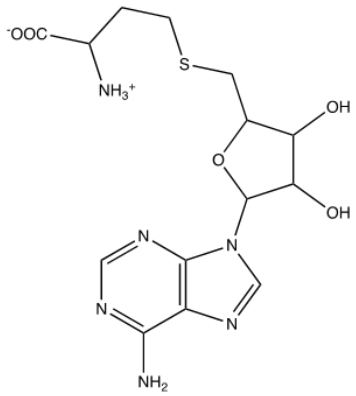


Fig. 5.4 Structure of S-adenosyl-L-homocysteine (SAH)

5.1.2 Field-Based Ligand Screener

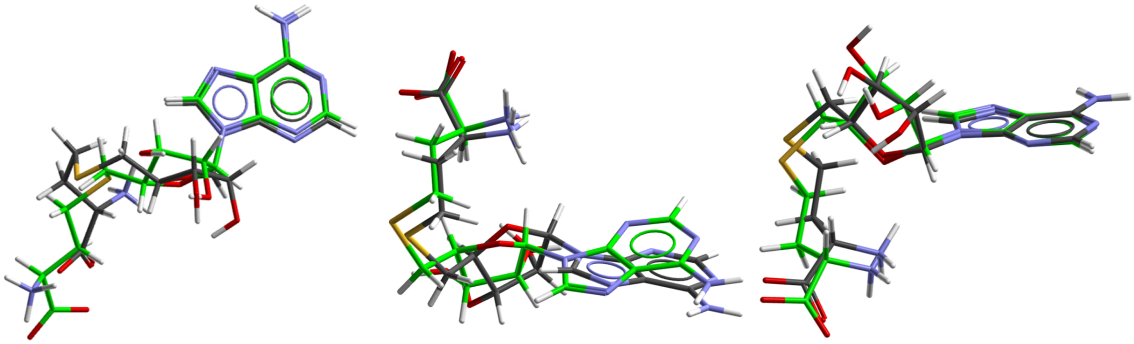
The field-based ligand screener is part of a larger ligand-based virtual screening workflow. In cases where a protein structure is absent, but multiple actives are known, it is possible to use this information to do 3D ligand-based virtual screening with the assumption that they bind to the same part of the protein. The workflow can be summarised as follows (see also, figure 5.6):

1. Generate 3D coordinates and enumerate conformers of active molecules.
2. Use the Ligand Overlay tool[184] to flexibly align common features of the conformers, aiming to generate a model for the bound position of the actives
3. Generate a field-based pharmacophore model to represent the common features of the actives. The Ligand Screener can then sample the fields with a library of molecules, scoring them with the values from the fields.

This section will explore the standard workflow for ligand-based virtual screening, and section 5.2.3 will use this context to explain how the Fragment Hotspot Maps can be inserted into this workflow.

5.1.2.1 Conformer Generation

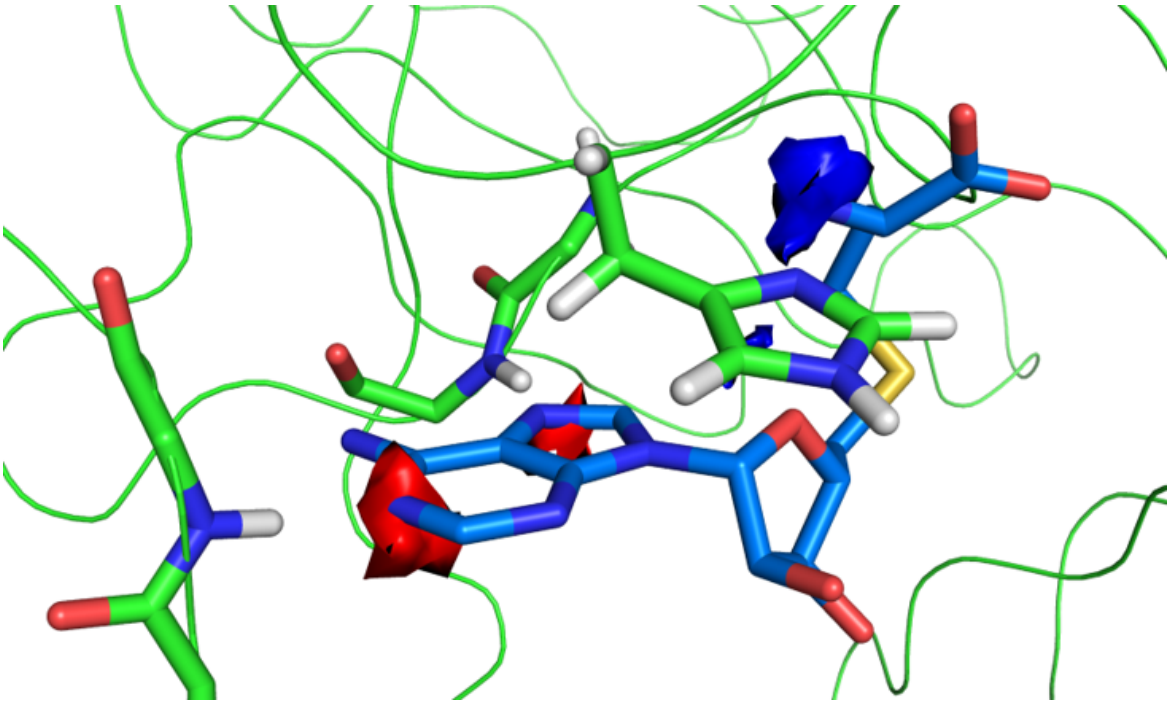
The CSD conformer generator takes an input 3D molecule, and makes use of structural data in the CSD to generate a realistic ensemble of low energy conformers. The CSD can be used to generate libraries of bond lengths, valence angles and rotamers[192]. These libraries are



(a) Default GOLD settings

(b) Extra sampling

(c) Hotspot-derived constraints



(d) Crystal structure with Fragment Hotspot Maps contoured at 17 (apolar hidden for clarity)

Fig. 5.5 Docking SAH into MLL1. Images (a), (b) and (c) taken from the GOLD tutorial, docked poses in green, crystal poses in grey.

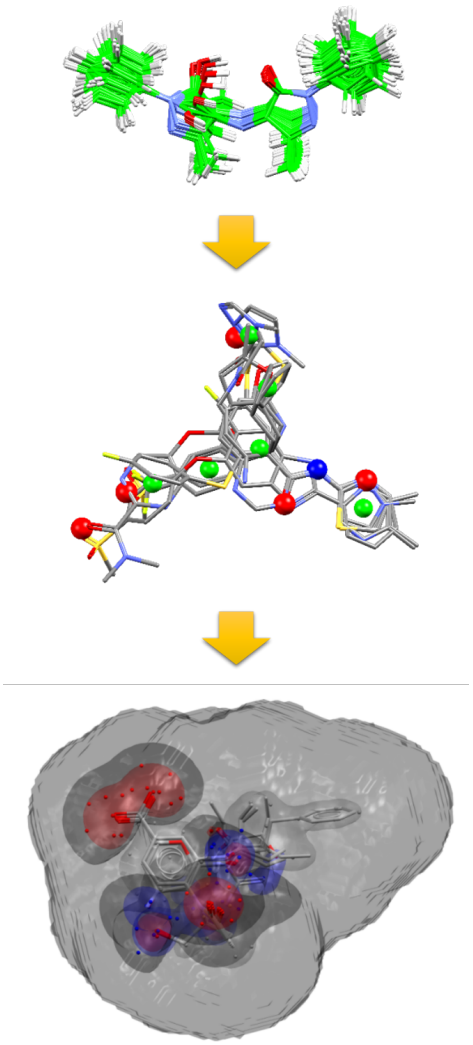


Fig. 5.6 Ligand-based virtual screening workflow. (Top) Conformer generation, (middle) ligand overlay and (bottom) field-based ligand screener. Adapted from the CCDC Ligand-based virtual screening tutorial

applied to a fragmented view of the input molecule and assigned an approximate probability score based on how frequently such a geometry is observed in the CSD. As a conformer is incrementally generated, it is checked for clashes and rejected as soon as a clash is detected. The conformers are clustered by conformer similarity, and a diverse set returned. The CSD conformer generator is available through the CSD Python API, and readily incorporated into greater workflows

5.1.2.2 Ligand Overlay

The ligand overlay application[184, 193] takes a set of active molecules, and aims to flexibly align them such that ligand groups that interact with a given protein residue are superimposed. In the absence of a protein structure, this model can give information about the binding site shape and key interactions.

Each ligand is annotated with features, such as hyrdophobe, strong donor, medium acceptor *etc*, which are described by a set customisable smiles arbitrary target specification (SMARTS) strings. Once the features have been identified, fitting points are placed either at the location of the heavy atoms or the centroid of a set of atoms, depending on the feature type. The overlays are represented as fingerprint called a chromosome, as described above for sampling with GOLD, and bit-string manipulations are applied to these fingerprints to generate thousands of overlays. These overlays are subsequently scored using three functions, union volume, hydrogen bond match, and hydrophobic match. They are ranked by constrained Pareto ranking, and the top twenty overlays of a diverse subset of the solutions are returned.

5.1.2.3 Ligand Screener

The ligand screener uses a set of overlaid molecules to build a grid-based pharmacophore model, which it then uses to screen a library of molecules. The overlaid molecules can come from the ligand overlay application described above, but can also come from a set of aligned protein-ligand crystal structures. Each atom is assigned a feature type (strong donor, medium donor, weak donor, strong acceptor, medium acceptor, weak acceptor, donor-acceptor and non-polar), much like in the ligand overlay program. Once all the features are assigned, a field potential is created for each feature type by placing a gaussian distribution at each atom location. The resulting grids reflect how frequently a given interaction type is found at each position, and larger sets of actives can give a better idea about which interactions are more

important for binding. In addition to this, an excluded volume penalty is also created for grid points greater than 3 Å away from any atoms in the overlay.

Fitting points are created from the model, and a global optimisation of the translation and rotation is performed to fit conformers of the library ligands. Customisable rules allow the user to weight the scoring of a match, for example between a "medium donor atom" from a screened ligand to a "strong donor fitting point", as well as penalties for mismatched placements. The best scoring pose it used for each ligand, and a ranked list returned.

5.1.3 Metrics

There are multiple ways to assess the performance of virtual screening methods. The results discussed within this chapter will use metrics suggested by Jain and Nicholls [194], receiver operating characteristic (ROC) curves in conjunction with area under curve (AUC) and enrichment factor (EF) calculations. ROC curves are a type of plot that is used to assess a binary classifier system, in this case the ability of a virtual screening method to classify molecules as active or inactive. An example ROC curve is shown in figure 5.7, and this particular example shows a method that is capable of retrieving active molecules more successfully than a random selection, demonstrated by the green curve being above the random line. The AUC of the random line is 0.5, therefore any value greater than 0.5 shows improvement.

One criticism of using AUC as a virtual screening method is that it is a global measure. In reality, the typical usage of a virtual screening workflow is to take the top x% of the ranked list to confirm with experimental screening. A successful virtual screening tool should enrich the top n% with active molecules, and the enrichment factor, described by equation 5.2, gives a measurement of this enrichment. $N_{experimental}$ describes the number of actives found after virtual screening, N_{active} is the total number of actives in the library and x% is the percentage of the library screened.

$$E_F = \frac{N_{experimental}}{N_{active} \cdot x\%} \tag{5.2}$$

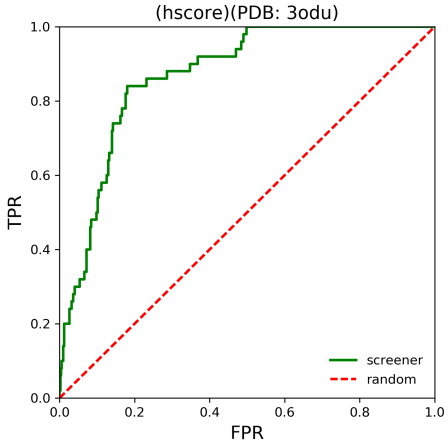


Fig. 5.7 Example ROC curve. The dotted line represents the performance of a random selection. False Positive Rate (FPR) is on the x axis, and True Positive Rate (TPR) on the y axis. While the green line is above the random line, the method is performing better than random

5.2 Method

In the rest of this chapter, we will assess how the scores from Fragment Hotspot Maps can be used to aid virtual screening. The first approach will assign a docking constraint within GOLD to the highest scoring polar interaction, placing a penalty on all ligand poses that do not interact with the selected protein atom.

5.2.1 Dataset

When evaluating a virtual screening method retrospectively, decoy molecules are screened as well as known actives, and the virtual screening method is assessed on its ability to rank the compounds such that the actives are on top. Extra special care must be taken when choosing active and decoy molecules. If the properties of the decoy molecules are different from the actives, the docking program may display artificially good performance. The directory of useful decoys - enhanced (DUD-e) set[195] provides a dataset of targets with diverse actives selected from ChEMBL, and property matched ligands from the ZINC database[196] as decoys. The full dataset contains 102 proteins, 22886 clustered active ligands from ChEMBL, each with 50 decoys from ZINC. As this work aims to be a preliminary look into the effect of using hotspots to guide virtual screening, the full set will not be used, but instead the diverse subset will be used (table 5.2). Each protein was protonated using protoss[164], with all waters and ligands removed (cofactors such as the heme group of CP3A4 are kept).

Table 5.2 Diverse subset of DUD-E

Target Name	PDB	Description	Actives
AKT1	3cqw	Serine/threonine-protein kinase AKT	293
AMPC	1l2s	Beta-lactamase	48
CP3A4	3nxu	Cytochrome P450 3A4	170
CXCR4	3odu	C-X-C chemokine receptor type 4	40
GCR	3bqd	Glucocorticoid receptor	258
HIVPR	1xl2	Human immunodeficiency virus type 1 protease	536
HIVRT	3lan	Human immunodeficiency virus type 1 reverse transcriptase	338
KIF11	3cjo	Kinesin-like protein 11	116

5.2.2 Hotspot-Guided Docking

Following the workflow described in the previous chapter's figure 4.5, the Hotspot API was used to calculate the Fragment Hotspot Maps, and the "predict protein H bond constraint" function was used to assign a constraint to the highest scoring polar interaction. The ChemPLP scoring function[197] was used, as it has previously been identified as the best scoring function for virtual screening[198] and is now the default scoring function in GOLD. For each ligand, 15 docks were permitted, and all other settings were left as the default. The docking calculations were performed twice for each protein, both with and without the protein H bond constraint. As this work aims to treat pockets as completely novel, only the automated optimisation of the hydrogen bonding network by Protoss was performed, and no other adjustments were made to the protein residues.

5.2.3 Hotspot Field-Based Screening

A crucial step in the standard ligand-based virtual screening workflow is the generation of field potentials to describe where ligand atoms are most often placed in the overlay. In the screening API, the only input allowed for the ligand screener is a set of overlaid ligands, and these field potential grids are only created as an intermediate step. Modifications made by Jason Cole, a colleague at the CCDC, allowed the calculations to be performed starting from the field potential grids. Although Fragment Hotspot Maps can be easily saved in the same file format as those required by the ligand screener, modifications needed to be made in order to use the Fragment Hotspot Maps in place of the normal field potentials:

Problem Polar interactions have three map types: strong, medium and weak

Solution Three grids for each polar probe are created. The strong grid contains grid points at the 95th percentile, the medium grid at the 60th percentile, and weak at the 30th percentile

Problem A donor_acceptor grid is required for atoms that can be both a donor and acceptor

Solution A donor_acceptor grid is created which includes all grid points where both the donor and acceptor score is greater than 14

Problem Favourable interactions are given negative scores, unfavourable are given positive

Solution The Fragment Hotspot Maps are multiplied by -1, and normalised such that the highest scoring grid point is -6, -3 or -1 depending on whether it is a strong, medium or weak grid. Non polar and donor_acceptor grids are treated as medium. The "strong" grids represent the highest scoring regions of the polar fragment hotspot map, and are weighted more favourably in the scoring step during screening. Unfavourable scores are not included, other than the excluded volume penalty

Problem Each grid needs an excluded volume penalty

Solution The ligsite grid is used to define which grid points clash with the protein. Any grid point with a ligsite score of 0 is given a score of 10 (unfavourable), otherwise the point is set to 0. A smoothing function is applied to give a gradient of scores from 0 to 10. The resulting excluded volume grid is added on top of all other grids

The modified grids (figure 5.8) are saved into a single directory with the correct file names, as required by the ligand screener tool. The final step is to provide conformers from the screening library. Using the CSD conformer generator, 25 conformers for each ligand were produced and screened. Previous unpublished worked showed that the using the default 200 conformers gives no improvement in performance. The values chosen above (percentiles and normalisation scores for strong, medium and donor) were selected as reasonable estimates for this proof of concept work, and have not been optimised. This approach will be referred to as the hotspot-based ligand screener (HS-LS) approach

In addition to using the Fragment Hotspot maps as input, the same experiment has been performed by Pete Curran, a fellow PhD student at the CCDC, using ligand bound protein crystal structures from the PDB. Using the crystallographic overlay represents the best case scenario, these results represent the best performance from the ligand screener tool regardless

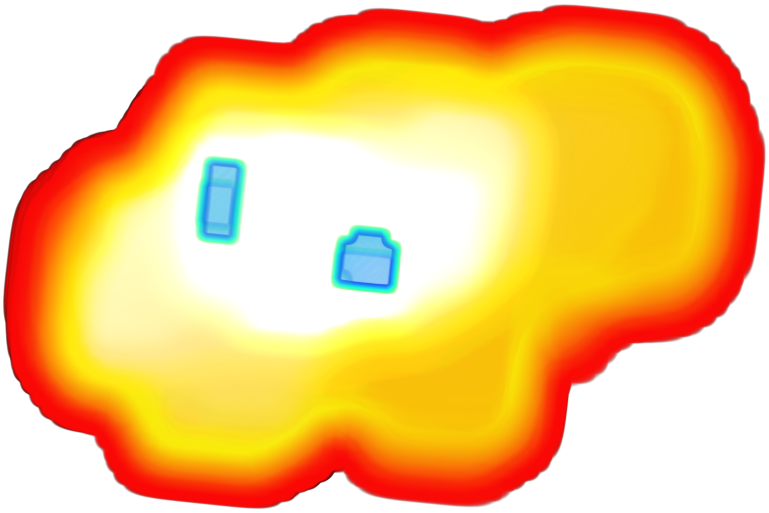


Fig. 5.8 A cross-section of a strong acceptor ligand screener grid generated from the acceptor fragment hotspot map. The colours represent favourable positions (blue) and unfavourable positions (red) for acceptor atoms.

of the quality of input. This method will be referred to as the PDB overlay ligand screener (PDB-LS) method. Curran performed a second run, having first removed crystal structure ligands that have a Tanimoto similarity ≥ 0.7 to the actives. This tests the ligand screener's ability to retrieve novel chemistry, and will be referred to as the novel PDB overlay ligand screener (nPDB-LS) method

5.3 Results and Discussion

Each virtual screening run for each target was run on a single processor. The time taken for each calculation varied based on the number of ligands to be screened, but the ligand screener was approximately 4 times faster than docking (9 hours vs 35 hours for KIF11). The following two sections will give an overview of the results for each of the methods, before taking a closer look at each target.

Table 5.3 Docking AUC

Target	Docking	Constraint	Change
akt1	0.78	0.79	+0.01
akt1 penalty=100	0.78	0.84	+0.06
ampc	0.52	0.57	+0.05
cp3a4	0.63	0.63	0
cxcr4	0.62	0.62	0
gcr	0.56	0.57	+0.01
hivpr	0.75	0.76	+0.01
hivrt	0.68	0.71	+0.03
kif11	0.81	n/a	n/a
Average	0.67	0.68	+0.01

5.3.1 Hotspot-Guided Docking Overview

The AUC values for docking with and without constraints are shown in table 5.3. It is clear from the table that there is only a modest improvement in AUC upon addition of the constraint. This might suggest that the default penalty for missing the selected protein hydrogen bond may have been too weak. In order to investigate whether a larger penalty would have further improved the results, AKT1 was run a second time with a penalty of 100. This resulted in a much larger improvement in both AUC and EF, demonstrated by the ROC curve in figure 5.9d. Due to the time required to run these calculations, it was not possible to repeat this for further targets.

As discussed previously, AUC is a global measure and EF more closely reflects the use of virtual screening in drug discovery. Enrichment at 1% is given in table 5.4, and shows improvement across all targets with non-zero enrichments. The two targets with $EF = 0$ have failed to find a single active in the top 1% of compounds screened, however these two targets have a much smaller number of decoys and ligands. As a result, the $EF_{1\%}$ is a less significant value for these targets, and $EF_{10\%}$ (table 5.5) is more suitable. $EF_{10\%}$ showed much smaller changers upon addition of the constraint, resulting in improvement for four targets and a reduction for two targets.

5.3.2 Hotspot Field-Based Screening Overview

AUCs for the three ligand screener experiments are shown in table 5.6. The "PDB-LS" column represents the best case scenario for the ligand screener, whilst the "nPDB-LS"

Table 5.4 Docking Enrichment at 1%

Target	Docking	Constraint	Change
akt1	9.90	12.73	+2.83
akt1 penalty=100	9.90	15.70	+5.80
ampc	0.00	0.00	0
cp3a4	4.12	7.06	+2.94
cxcr4	0.00	0.00	0
gcr	5.04	8.52	+3.48
hivpr	12.50	15.30	+2.80
hivrt	5.33	7.10	+1.78
kif11	17.24	n/a	n/a
Average	6.77	8.71	+1.94

Table 5.5 Docking Enrichment at 10%

Target	Docking	Constraint	
akt1	5.32	5.10	-0.22
akt1 penalty=100	5.32	5.94	+0.62
ampc	1.46	1.25	-0.21
cp3a4	2.65	2.82	+0.17
cxcr4	0.75	1.00	+0.25
gcr	3.14	3.14	0
hivpr	4.55	4.72	+0.17
hivrt	2.63	3.08	+0.45
kif11	6.21	n/a	n/a
Average	3.34	3.42	+0.08

Table 5.6 Ligand Screener AUC values. The "HS-LS" column has used modified Fragment Hotspot Maps as input, the "PDB-LS" column has used an experimental overlay as input, and "nPDB-LS" has used a subset of PDB ligands dissimilar from the actives in the test set. Cells with n/a show cases where no PDB ligands remained after the similarity cut off was applied

Target	HS-LS	PDB-LS	nPDB-LS
akt1	0.72	0.72	0.72
ampc	0.52	0.76	0.76
cp3a4	0.57	0.60	0.52
cxcr4	0.70	0.72	n/a
gcr	0.82	0.87	0.73
hivpr	0.79	0.87	n/a
hivrt	0.50	—-	0.50
kif11	0.81	0.91	0.91

column reflects the ability to retrieve novel chemistry. The HS-LS results tend to yield similar AUC values to the PDB overlay, and in some cases have better AUC values than nPDB-LS results.

Enrichment factors at 1% show varying performance between targets. There are large enrichments from both PBD overlay runs for gcr, hivpr and kif11, but these are not achieved by the HS-LS runs. Closer inspection of the ligand screener grids created from the PDB overlays showed that there was a much larger range of scores than the normalised scores created from the fragment hotspot maps. While the hotspot grids were normalised to be in the range of -6 to 0, the PDB overlay for kif11, the target with the best EF, resulted in a score range of -29 to 98 for the strong acceptor and -25 to 90 for the strong donor grids. In contrast, the worst performing PDB-LS run was cp3a4, which had a score range of -9 to 16 for the strong acceptor and -4 to 17 for the strong donor grid. In order to achieve the EFs of the PDB overlay methods with the hotspot input, it is likely that a greater score separation is required between the interactions found at the 95^{th} percentile, 60^{th} percentile and 30^{th} percentile. It may also be necessary to implement penalties for mismatched hydrogen bonding groups. This would require identifying grid points with a high score for one polar probe, but a low score for the other. As a result, favourable regions on the donor maps would appear as penalties on the acceptor maps, and vice-versa.

Table 5.7 Ligand screener Enrichment at 1%

Target	HS-LS	PDB-LS	nPDB-LS
akt1	7.09	7.09	0.71
ampc	0.00	8.06	4.84
cp3a4	2.75	2.20	2.48
cxcr4	0.82	2.46	n/a
gcr	3.55	21.85	17.76
hivpr	8.67	21.00	n/a
hivrt	1.10	—-	2.03
kif11	5.08	25.89	20.30

Table 5.8 Ligand screener enrichment at 10%

Target	HS-LS	PDB-LS	nPDB-LS
akt1	3.10	2.55	2.65
ampc	1.46	4.52	3.39
cp3a4	1.79	1.82	1.07
cxcr4	2.05	2.38	n/a
gcr	3.80	5.68	3.23
hivpr	4.14	5.87	n/a
hivrt	1.19	—-	1.33
kif11	3.55	6.50	6.45

5.3.3 AKT1

Serine/threonine-protein kinase AKT (AKT1) gave the second best AUC and EF for docking, and the fourth best AUC (second for EF) for the ligand screener. Introduction of the default constraint for docking gave a moderate improvement in AUC, but did see a larger improvement in EF $_{1\%}$ (9.9 to 12.7). The resulting ROC curves for with (figure 5.9b) and without (figure 5.9a) the hotspot derived constraint show little difference, however increasing the penalty to 100 gives a noticeably steeper curve. Inspection of the Fragment Hotspot Maps and the selected protein atom for constraint, shown in figure 5.9e, shows that the selected interaction is the backbone NH of the kinase's hinge.

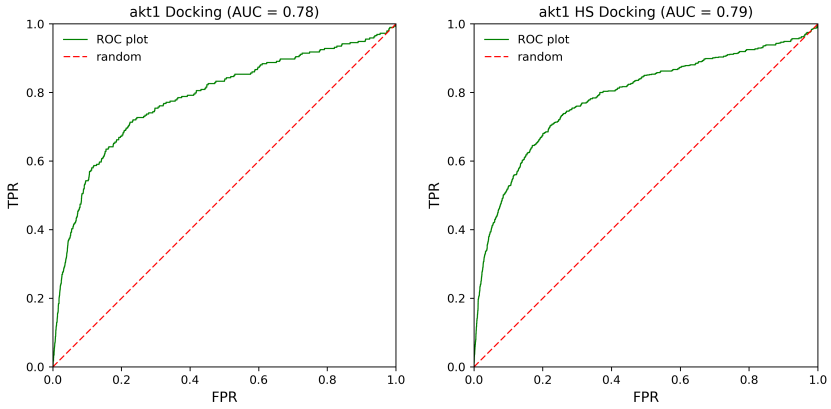
The ligand screener achieved both a slightly lower AUC and EF $_{1\%}$. Although the AUC was the same for all three ligand screener inputs, removal of PDB ligands similar to those in the active set led to loss of enrichment at 1%. This suggests that using the ligand screener with Fragment Hotspot Maps as input is better at discovering novel chemistry than the PDB overlay in this case.

5.3.4 AMPC

Beta-lactamase (AMPC) was one of the most challenging targets for either of the virtual screening methods. They offered almost no improvement over random selection for docking (figure 5.10a) or the ligand screener figure 5.10c. Addition of the hydrogen bond constraint to the amino group of asparagine 152's carboxamide leads to improvement in AUC from 0.51 to 0.57. Despite the improvement in AUC, this still results in an EF of 0, meaning no actives were found in the top 1%. This is not too surprising, as AMPC only has 48 actives. Using equation 5.2, a single active found in the top 1% would give an EF of 2.08. For so few actives, it is more meaningful to take the EF at 10%, which in this case gives 1.46 for normal docking, 1.25 with the constraint.

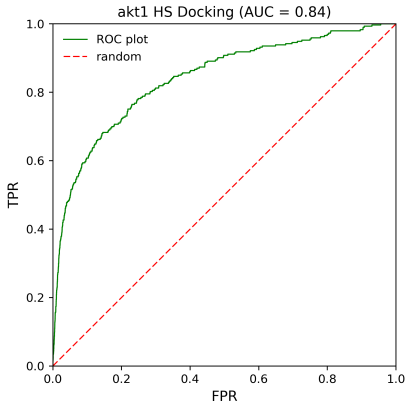
The results from the ligand screener using the PDB overlay show much better results both with and without the similar molecules removed, as compared to the other methods. Visual inspection of the overlaid PDB ligands (figure 5.11) show that although most of them do make the highlighted interaction, they extend towards the left, whereas the apolar map used for the ligand screener highlights the more buried region to the right. It is unlikely that the ligand screener will attempt to place the ligands in the correct orientation as a result of this.

This binding site yields very low scoring Fragment Hotspot Maps. The maps in figure 5.11 were contoured at an absolute score of 12, showing that this site does not contain a

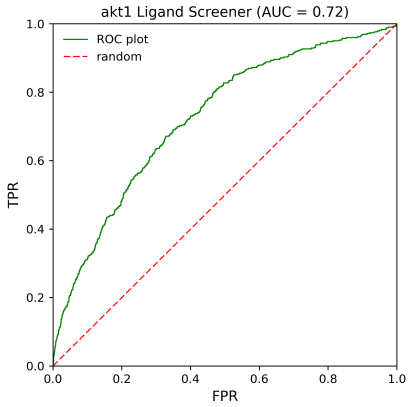


(a) Docking

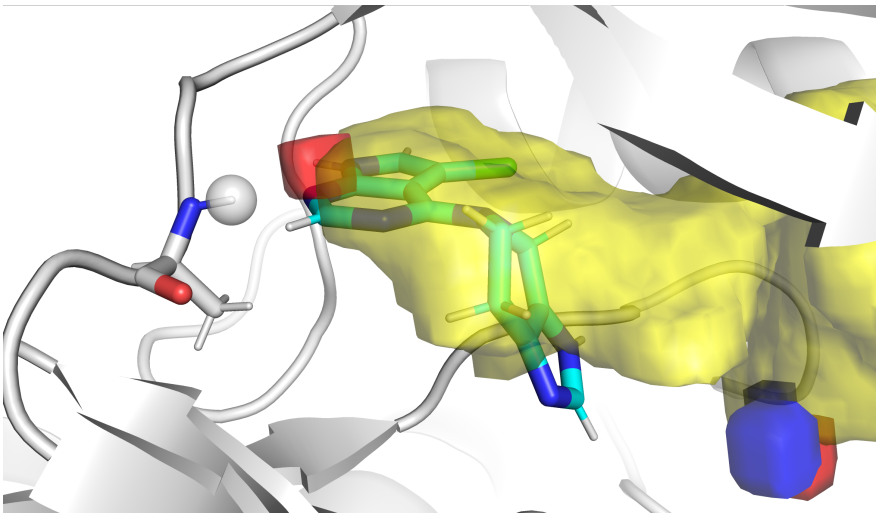
(b) Docking with constraint



(c) Docking with strong constraint



(d) Ligand Screener



(e) Crystal ligand displayed with maps contoured at the 95th percentile, an additional contour level of 17 for the apolar map is also displayed. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.9 ROC curves and binding site for AKT1

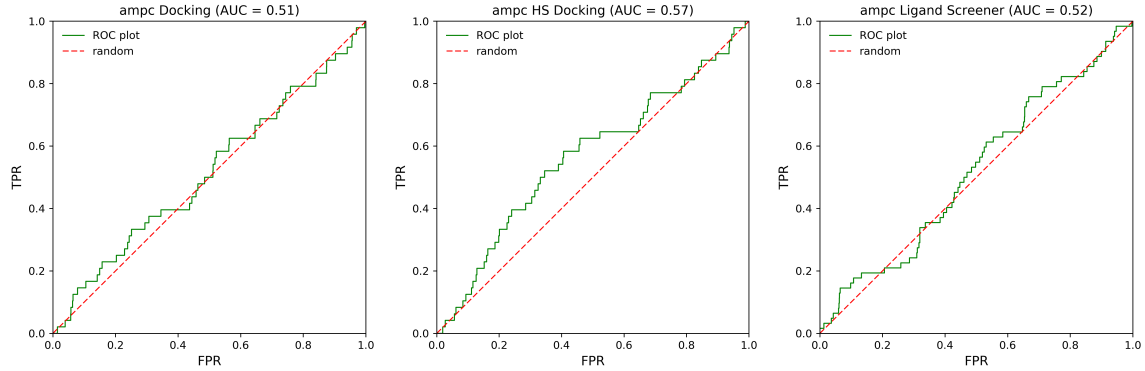
hotspot. Fortuitously, a ligand deconstruction experiment was performed on this target by Babaoglu and Shoichet[199]. Upon deconstruction of this ligand into fragments, none of the fragments managed to recapitulate the binding position of the larger ligand. Previous work [200] by Kozakov and co-workers found that fragments from a deconstructed ligand would only maintain its binding position if it had sufficient overlap with a hotspot. Typically, fragmentation would result in multiple fragments occupying a single hotspot, however the work by Babaoglu and Schoichet found that the fragments were located in several new binding positions. This indicates that a hotspot is not present.

Interestingly, while AMPC is normally rigid and undergoes little conformational change upon binding of lead-like molecules, the small fragments do result in conformational change. Maps calculated from the fragment bound structures (PDB: 2HDQ) show a strong hotspot present (figure 5.12). This suggests that the dynamic motion of the protein generates conformations which create a hotspot that larger ligands are unable to satisfy, but that a fragment is capable of stabilising.

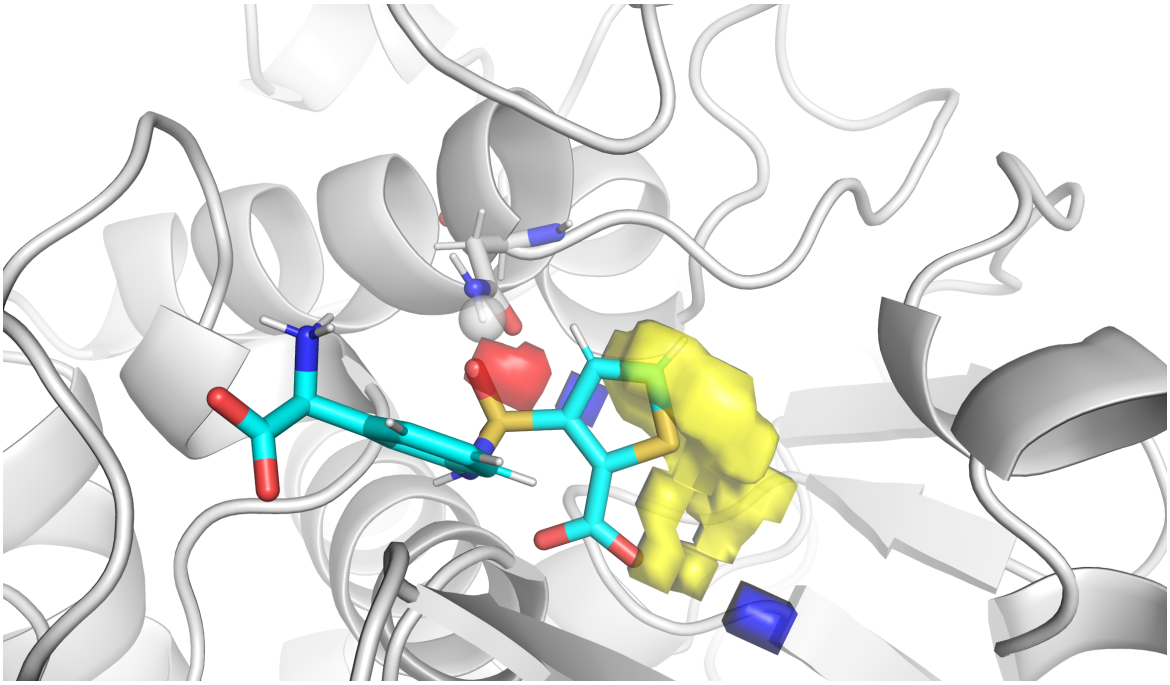
5.3.5 CP3A4

Cytochrome P450 3A4 (CP3A4) is a metabolic enzyme found in the human liver, and known to bind a large variety of drug-like molecules[201]. As a result, it is not expected to be a particularly good candidate for virtual screening. AUCs for docking are again fairly low, but showing improvement over random. An AUC of 0.63 was achieved with or without the constraint included (figures 5.13a and 5.13b).

The constraint was applied to the hydrogen of the serine 119's hydroxyl group (figure 5.13d), meaning that the hydroxyl was acting as a hydrogen bond donor. SuperStar, and by extension Fragment Hotspot Maps, treat hydroxyls as rotatable, therefore there are overlapping regions of donor and acceptor propensity surround the serine. GOLD is also capable of rotating hydroxyl groups, however the constraint was applied to the hydrogen atom, and therefore not satisfied if the hydroxyl group was acting as an acceptor. The crystal structure ligand forms a hydrogen bond between the NH of the carbamate and the oxygen of the hydroxyl, however this results in the hydroxyl hydrogen pointing towards hydrophobic residues. Addition of the constraint manages to improve enrichment at 1% from approximately 4 to 7, suggesting that some actives do use the hydroxyl as a donor. It is possible that having the ability to act as donor or acceptor within a hotspot at CP3A4 contributes to its promiscuity.



(a) Docking **(b)** Docking with constraint **(c)** Ligand Screener



(d) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks.

Fig. 5.10 ROC curves and binding site for AMPC

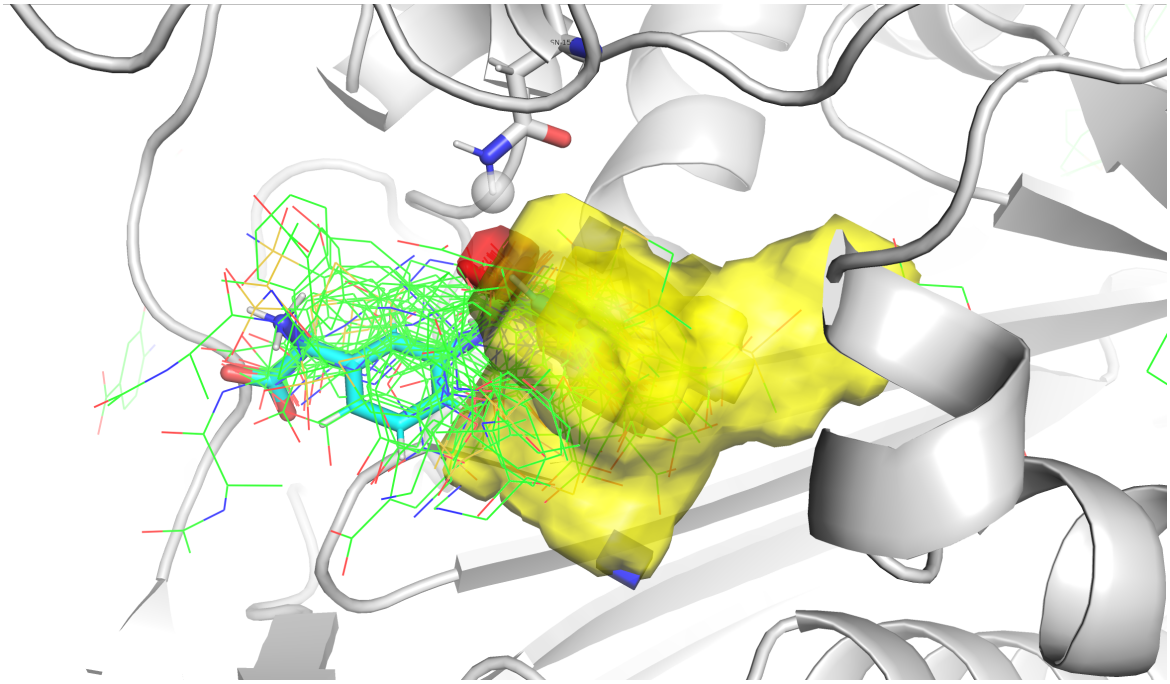


Fig. 5.11 AMPC with ligand overlay. Overlaid ligands are displayed as green lines, the crystallographic ligand is shown as cyan sticks. An additional contour of the apolar map is shown at the 60^{th} percentile, showing the minimal overlap between grid points used for the apolar ligand screener maps and the experimental ligands.

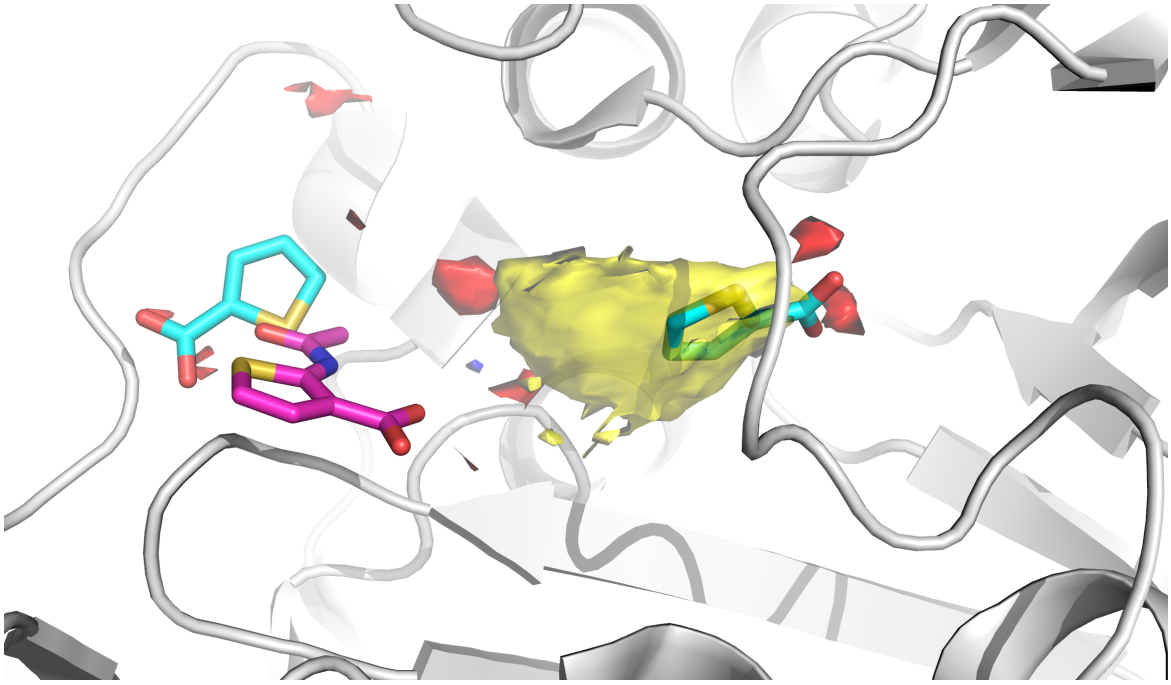
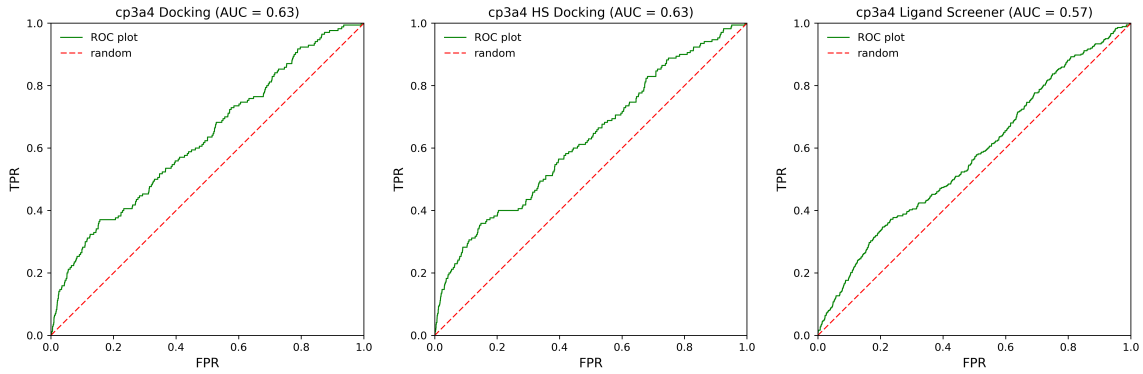
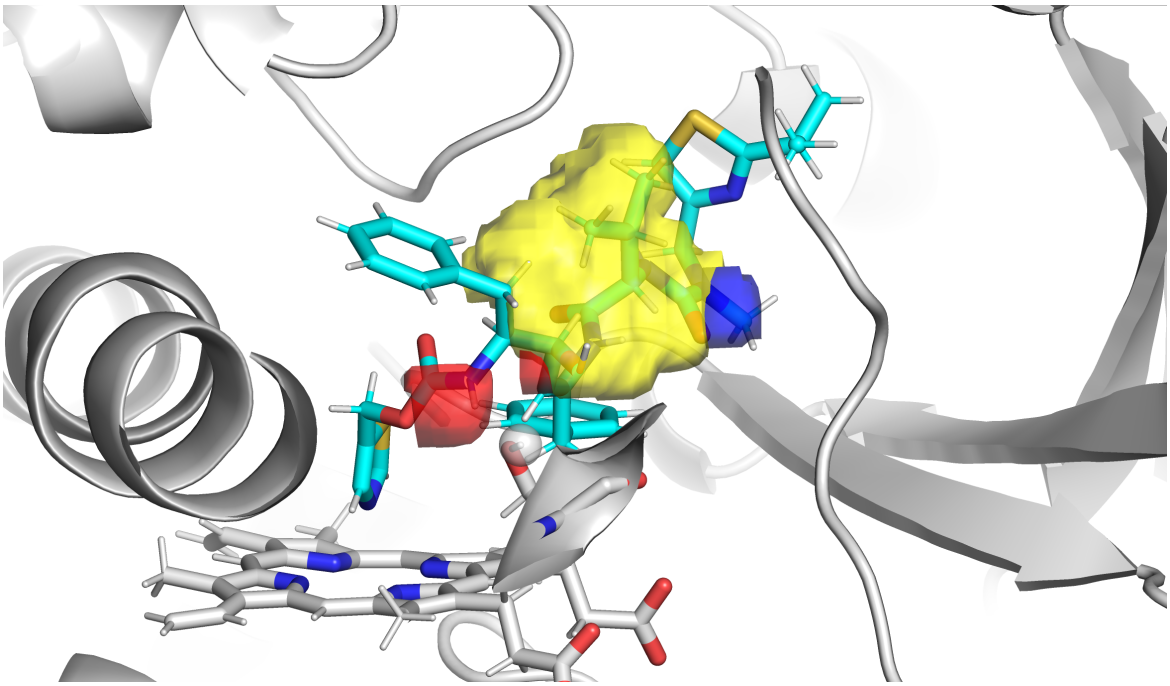


Fig. 5.12 AMPC with fragments. Small fragments with new binding mode are shown in cyan sticks. The larger fragment, which maintains its binding mode from its parent ligand is shown in magenta. Maps are calculated from the protein structure of the cyan fragments, apolar map is contoured at 17, acceptor map is contoured at 14.

The PDB-LS only just outperforms using the HS-LS, and removing PDB structures that are similar to those found in the active set (nPDB-LS) essentially reduces the performance down to random.



(a) Docking **(b)** Docking with constraint **(c)** Ligand Screener



(d) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks.

Fig. 5.13 ROC curves and binding site for CP3A4

5.3.6 CXCR4

C-X-C chemokine receptor type 4 (CXCR4) is an example where the ligand screener outperforms docking, however neither of the methods manage to achieve good early enrichment.

CXCR4 has the fewest actives in the dataset, and it is again more appropriate to look at the EF at 10%.

The polar atom selected for the docking constraint is one of the carboxylate oxygens of aspartate 95, which can be seen interacting with one of the nitrogens of the symmetrical isothiourea group. These are both protonated with a net positive resonance charge[202]. This interaction is important for affinity, and methylation of this nitrogen results in a 100-fold loss in potency[203]. Despite a suitable interaction being set as a constraint, it has minimal impact on either the AUC or EF at 10%.

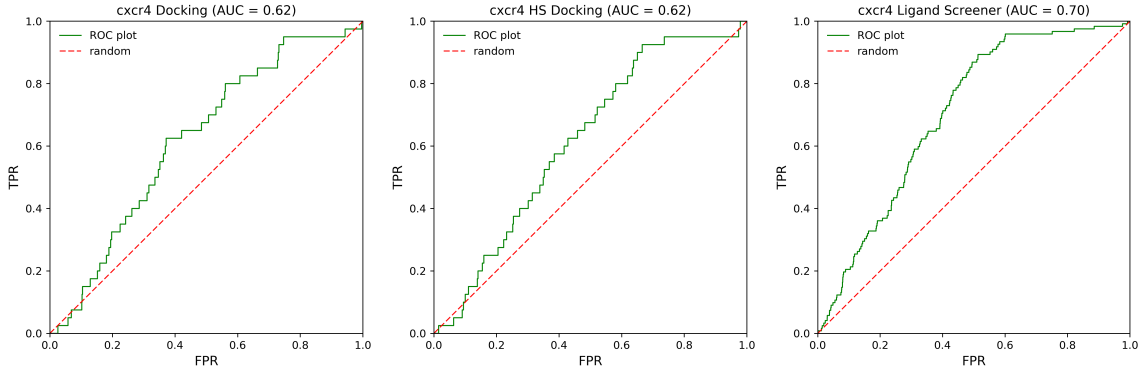
The HS-LS method managed to almost match both the AUC and EF at 10% of the PDB-LS method (tables 5.6 and 5.8. All of the PDB ligands were found to be above the similarity threshold to the active set, therefore it was not possible to calculate the nPDB-LS results.

5.3.7 GCR

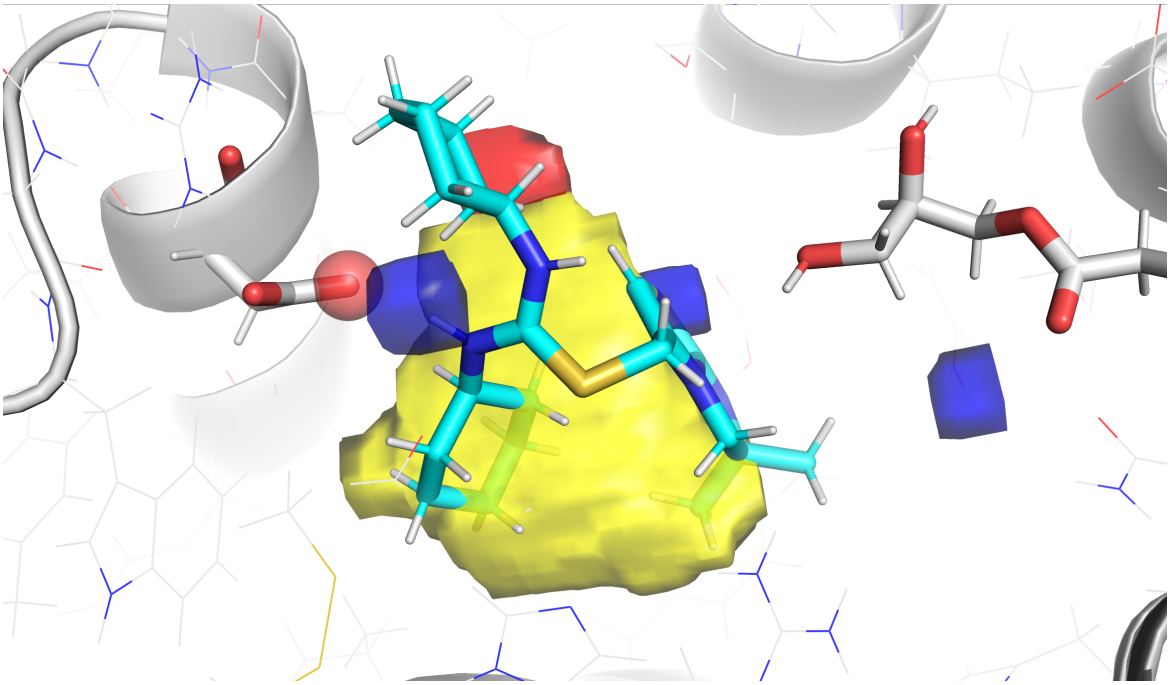
Glucocorticoid receptor (GCR) shows the greatest difference between docking and the ligand screener. Docking performance is fairly poor with respect to AUC, with AUCs of 0.56 (no constraint, figure 5.14a) and 0.57 (with constraint, figure 5.14b. Docking does give reasonable early enrichments (table 5.4), with 5.04 being improved to 8.52 upon addition of the constraint shown in figure 5.14d.

In the DUD-e publication[195] all targets were tested with DOCK 3.6, and GCR was in the bottom 4 of all 102 targets. It was described as a particularly difficult target to dock into as it was a very hydrophobic and flexible pocket. In contrast to the performance of docking, the ligand screener does very well. Where the ROC curves for the docking results (figure 5.14a) dip below the random line after initial enrichment, the ligand screener (figure 5.14c) stays well above it, resulting in much higher AUC values.

The difference in performance between docking and the field-based ligand screener for GCR suggests that the two methods may have orthogonal uses. It could be that the "fuzzy" nature of the ligand screener fields, whether generated from Fragment Hotspot Maps or a ligand overlay, are more suitable for flexible targets. Furthermore, the ability of Fragment Hotspot Maps to precisely locate the hydrophobic hotspot within the binding site allows for good performance with hydrophobic binding sites.

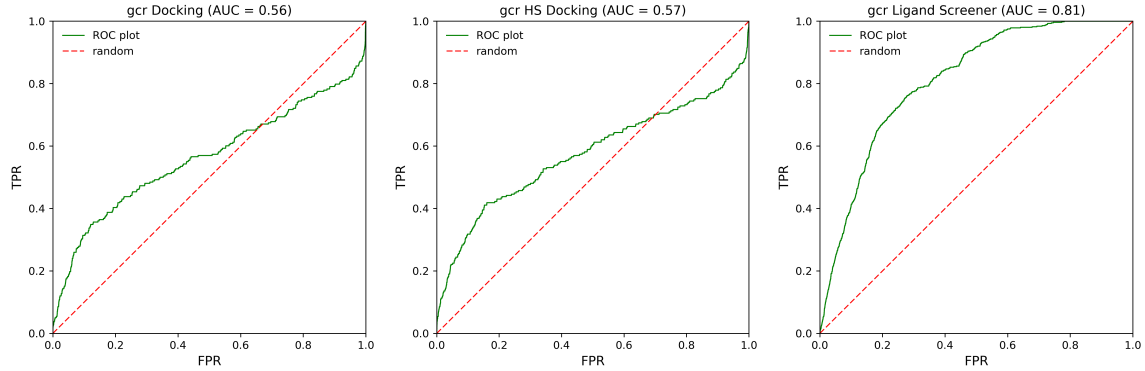


(e) Docking **(f)** Docking with constraint **(g)** Ligand Screener

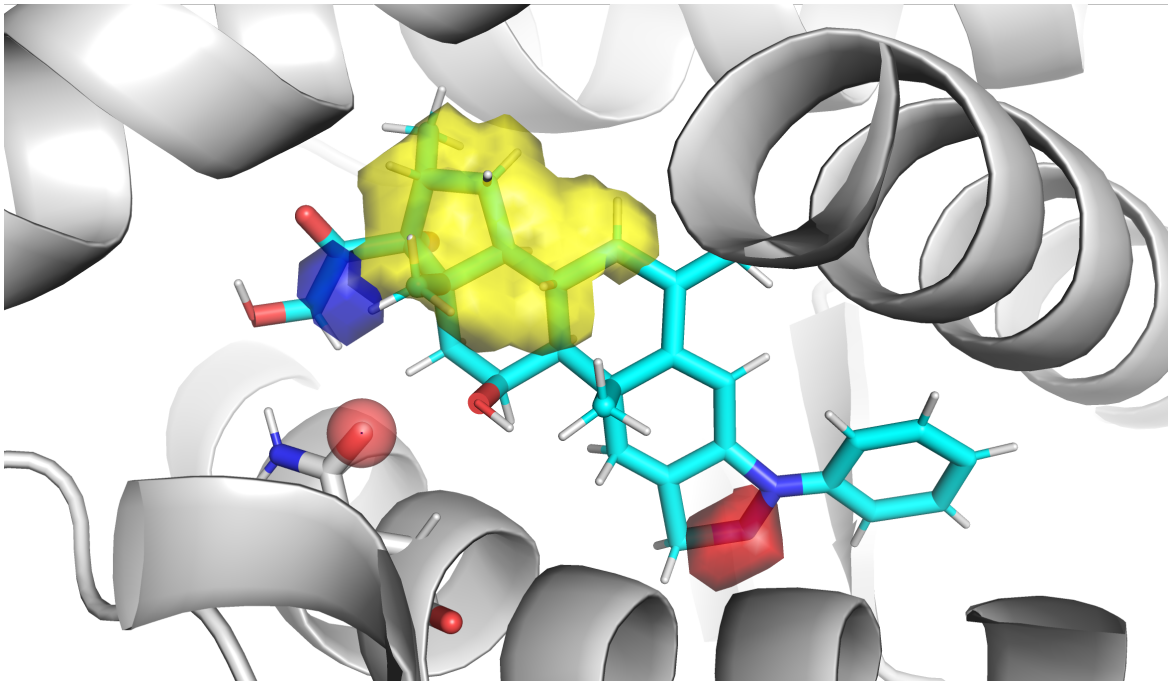


(h) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.13 ROC curves and binding site for CXCR4



(a) Docking **(b)** Docking with constraint **(c)** Ligand Screener



(d) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.14 ROC curves and binding site for GCR

5.3.8 HIVPR

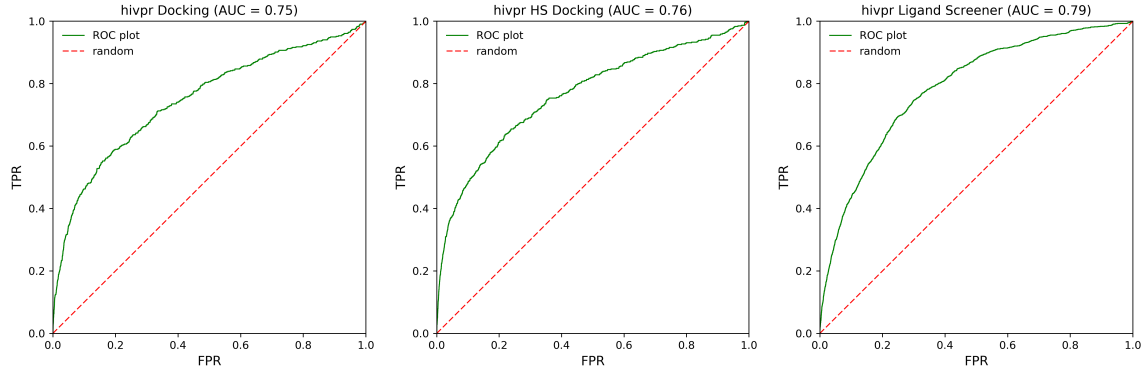
Human immunodeficiency virus type 1 protease (HIVPR) has a good AUC and early enrichment for both methods. In terms of AUC the ligand screener outperforms docking both with and without constraints, however docking gives a superior early enrichment. Both AUC and EF are improved for docking upon addition of the constraint. The aspartates of this aspartyl protease were not selected as the highest scoring interaction by the Fragment Hotspot Maps, as charged probes have currently not been implemented. Instead, the backbone carbonyl oxygen of Gly-34 was set as the protein hydrogen bond constraint, still resulting in an increase in early enrichment from 12.50 to 15.30.

5.3.9 HIVRT

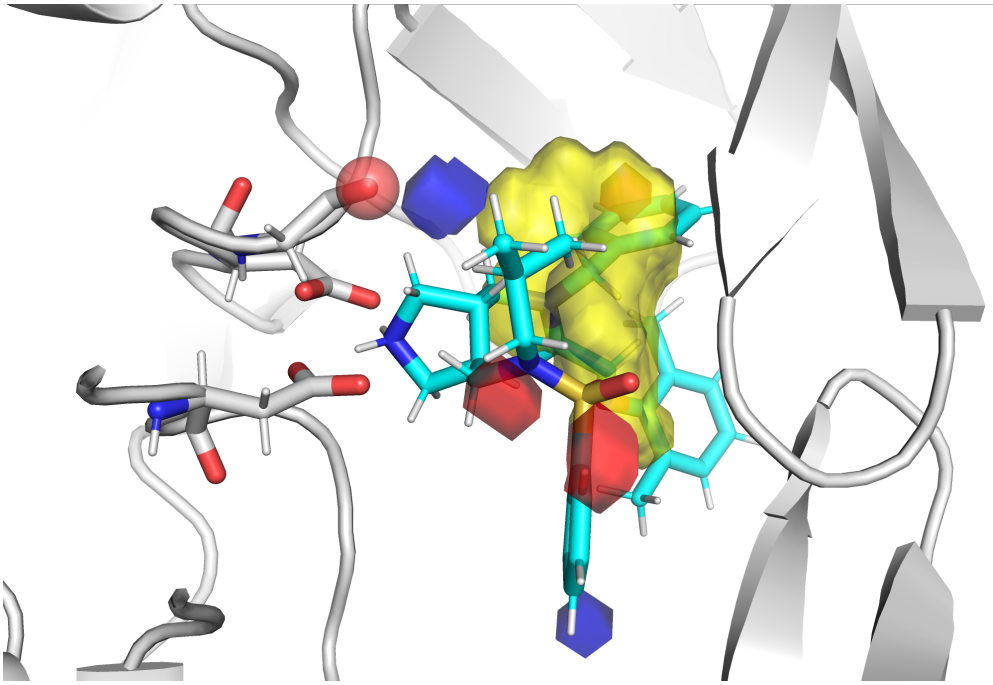
Human immunodeficiency virus type 1 reverse transcriptase (HIVRT) shows reasonable docking performance for both AUC and EF, but performs poorly with the the field-based ligand screener. Addition of the docking constraint improves both AUC and EF; the atom selected for the constraint is the only polar atom to interact with the crystallographic ligand (figure 5.16d). The poor performance of the ligand screener is surprising as the maps seem to match the crystallographic ligand well. Conformational change of the binding site is unlikely to be the cause of the poor performance, as this would have affected the docking results as well. Comparing the HS-LS results to the nPDB-LS results show that the experimental overlay of ligands performs equally badly.

5.3.10 KIF11

In this dataset, the binding site on Kinesin-like protein 11 (KIF11) is allosteric. From figure 5.17c you can see that the binding site is shown to be hydrohphobic and lacking in polar interactions. This is in line with an analysis of ligands in ChEMBL[204], which found allosteric inhibitors to be more lipophilic and rigid. As a result, no polar interactions were found above the score threshold required to define a protein hydrogen bond constraint. Despite the lack of polar features, KIF11 performed very well with both docking and the ligand screener. Comparing the different ligand screener approaches, KIF11 gave the second highest AUC values for the HS-LS method, beaten only by the other hydrophobic binding site of GCR. The AUC and EF values for the PDB-LS and nPDB-LS methods are the best for all targets across all methods.

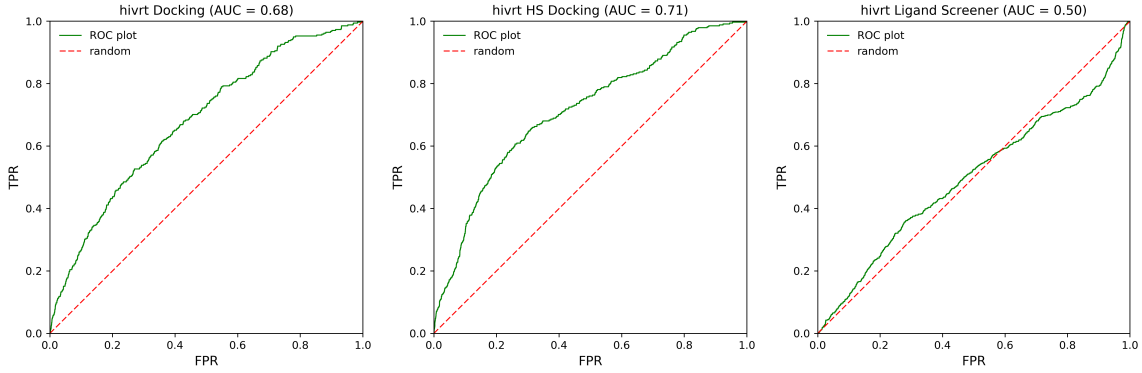


(a) Docking **(b)** Docking with constraint **(c)** Ligand Screener

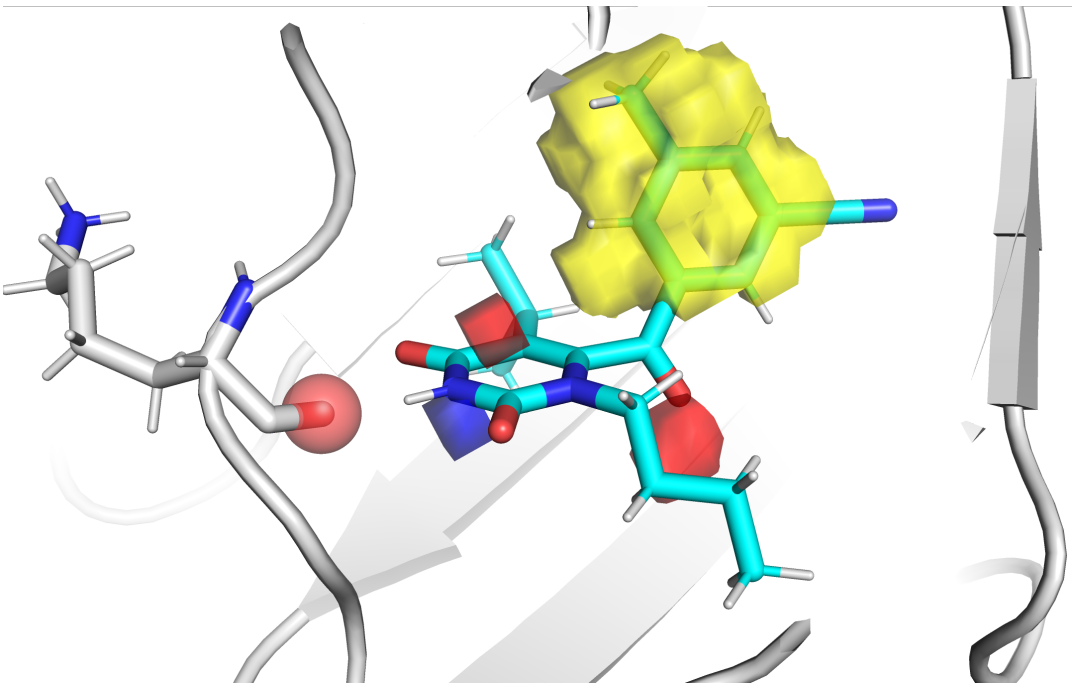


(d) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.15 ROC curves and binding site for HIVPR

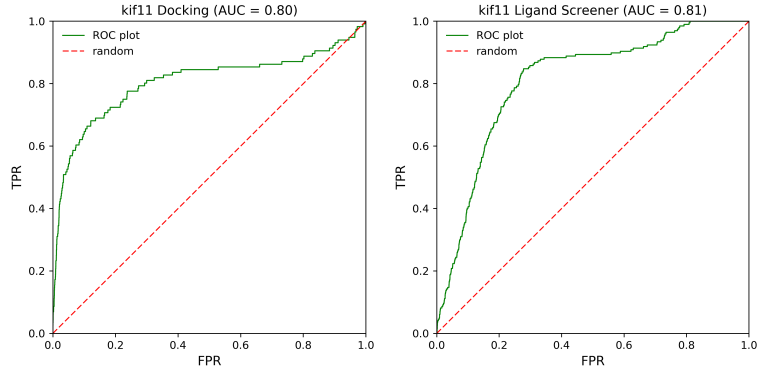


(a) Docking **(b)** Docking with constraint **(c)** Ligand Screener

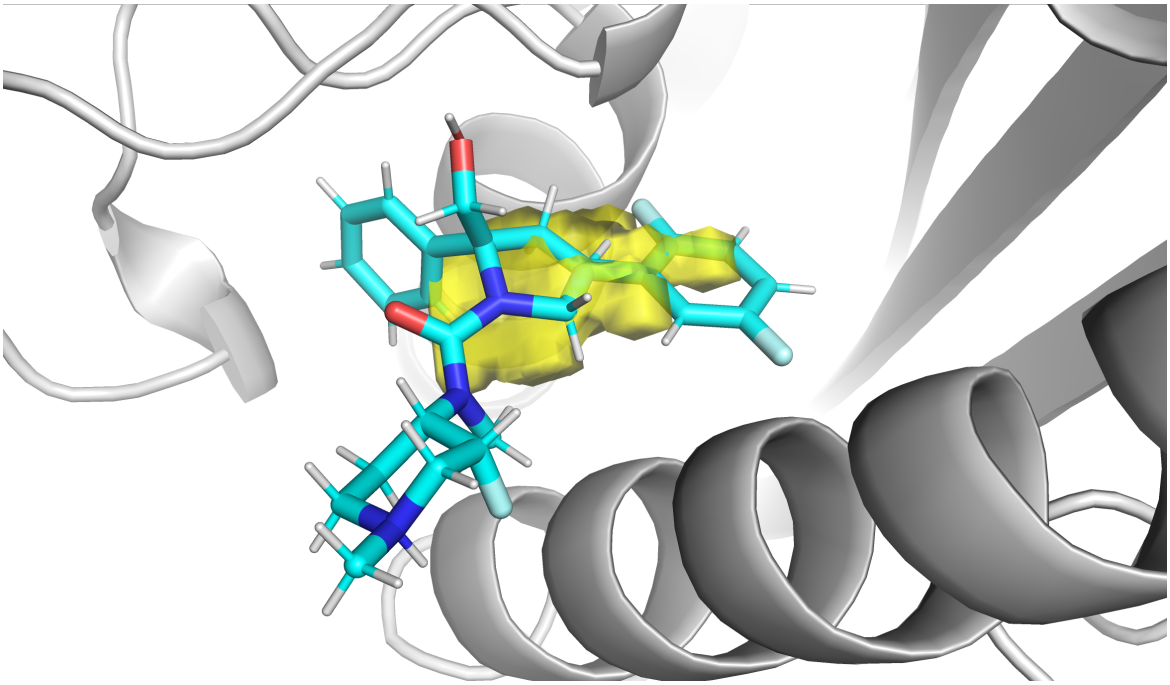


(d) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.16 ROC curves and binding site for HIVRT



(a) Docking **(b)** Ligand Screener



(c) Crystal ligand displayed with maps contoured at the 95th percentile. The sphere denotes the atom used as the hydrogen bond constraint in docking. Crystallographic ligand is shown as cyan sticks

Fig. 5.17 ROC curves and binding site for KIF11

5.4 Conclusion

The work carried out in this chapter aimed to use the knowledge gained from a Fragment Hotspot Maps calculation, and apply it to existing virtual screening tools. Using the Hotspot API functions discussed in chapter 4, two approaches were taken. Firstly, the highest scoring polar interaction in the binding site was set as a "protein hydrogen bond constraint", meaning that any pose that fails to make that interaction will have a penalty applied to it.

The second approach was to use the maps more directly, making use of the field-based ligand screener. This tool is typically part of a larger ligand-based workflow, however the final step involves sampling ligands in a field-based pharmacophore model. Fragment Hotspot Maps were adapted such that they can be used by the ligand screener program to screen molecules against the maps directly.

Docking constraints selected by choosing the highest scoring polar interaction in the binding site led to improved performance in all but two cases, where no change was observed. Improvements were small in terms of AUC but larger for EF. Increasing the penalty of the constraint from the default of 10 to 100 saw a much greater improvement in performance. This initial work has shown that information from Fragment Hotspot Maps can be used to improve the performance of docking, but further work is required to find the optimal penalty to use. While this work used a single polar constraint, it is also possible to select multiple constraints. Although not tested here, functionality has been added to the Hotspots API to convert the apolar maps to to hydrophobic fitting points. By default, GOLD generates hydrophobic fitting points by calculating the Van der Waals interaction between the protein and a carbon atom, and the Fragment Hotspot Maps method provides an alternative method for highlighting favourably hydrophobic areas. The default fitting points often fill much of the cavity, therefore those derived from Fragment Hotspot Maps offer an opportunity to have more targeted hydrophobic interactions.

Using the field-based ligand screener with modified Fragment Hotspot maps provides an alternative structure-based virtual screening workflow to docking. The performance of this approach was compared to using the field-based ligand screener with an input of overlaid ligands from multiple ligand-bound protein crystal structures. This represents the best case scenario, and shows the performance of the field-based ligand screener itself. Additionally, the ligand screener was provided a subset of these overlaid ligands with molecules similar to actives within the test set removed. This represents the field-based ligand screener's ability to find novel chemistry.

Comparing AUCs of calculations from docking and the field-based ligand screener shows that for most targets they give a similar performance, with the clear exception of GCR. The flexible and hydrophobic nature of GCR has made it a difficult target for virtual screening, reflected in the poor docking performance shown in this chapter. The good performance of the field-based ligand screener can potentially be attributed to two things. Firstly, the "softer" nature of field-based ligand screening can be more forgiving of clashes when the shape of the crystal structure does not match the bound conformation of the protein. This could make the field-based ligand screener a more suitable structure-based virtual screening tool for highly flexible sites and homology models. Secondly, the hydrophobic maps have be shown previously[1] to perform well at specifically locating the fragment binding site, allowing for more precise placement of hydrophobic groups in lead-like molecules.

Comparing the AUCs of the field-based ligand screener results from the three different input types shows that in cases where a reasonably strong hotspot is found, the HS-LS calculations can offer similar performance to the best case scenario of the PDB-LS. For all targets, percentiles of the map scores were used rather than the absolute values (95^{th} and 60^{th} rather than 17 and 14). This produced ligand screener maps with consistent volumes for each binding site, however AMPC did not have a predicted hotspot, and performed very poorly. When ligands similar to those in the test set are removed, nPDB-LS, AUCs were greater than those from HS-LS in only two out six applicable cases (two the same, two lower). This suggests that, in terms of AUC, virtual screening using the HS-LS method can offer similar performance at identifying novel chemistry to using experimentally overlaid ligands from protein-ligand crystal structures. The HS-LS method generally showed lower EFs for most targets when compared to both docking and PDB-LS. The low early enrichment shows a need for optimisation of the scores of the ligand screener grids generated from the Fragment Hotspot Maps.

The two methods discussed in this chapter provide orthogonal structure-based virtual screening approaches. While this work requires optimisation, it suggests that Fragment Hotspot Maps can be used in virtual screening both directly with the field-based ligand screener, and indirectly with constraints in docking. When combined with the Hotspots API, discussed in the previous chapter, these improvements can be achieved with simple Python scripts that require minimal knowledge of Python. Importantly, these approaches have demonstrated that they can improve virtual screening from the structure alone, making use of tools that typically require large amounts of experimental knowledge of the binding

site. The Hotspot API provides a framework for systematic detection of ligandable pockets in the PDB, and subsequently running a virtual screen against these pockets.

Chapter 6

Current and Future Uses of Fragment Hotspot Maps

6.1 Introduction

This chapter will explore recent work and examples of Fragment Hotspot Maps being used to learn about a target at varying stages of the drug discovery process, still a matter of ongoing research for me and collaborators. Each of these research areas will be discussed below. I will indicate what has been achieved so far and what plans there are for the future.

6.2 Pocket Tractability Assessment

Fragment hotspot maps can be used as a direct score of ligandability. As fragment hit rate has been shown to correlate with ligandability [25], a computational method that can predict fragment binding may also be able to predict ligandability. This has been demonstrated recently [110] by using hotspots predicted by FTMap, combined with the volume of the pocket and the density of hotspots. As the scores from my Fragment Hotspot Maps method are comparable across targets, due to their probabilistic origin of sampling SuperStar propensities, the cut-offs of 14 and 17 can be used to assess the ligandability of a target. This is ligandability in the true sense, as it is possible that a pocket is large enough to accommodate a hotspot, but not a drug-like molecule.

Since the start of August 2017 I have been working at the European Bioinformatics Institute (EBI) looking large scale tractability assessment as part of the open target platform.

As part of this, I have been investigating how Fragment Hotspot Maps can be used to aid large scale tractability assessment.

Structure-based tractability-prediction methods are sensitive to the initial pocket definition. Slight changes in the conformation of the binding site can cause separate pockets to be considered as one, yielding extremely large pockets. This causes problems for methods that have been trained on cavity properties, as the global measures used change drastically. Volkamer and colleagues attempt to deal with this problem with DoGSiteScorer, by using the difference-of-Gaussians method to detect subpockets [37]. These subpockets are ultimately joined together to create pockets, however properties and druggability scores are created for both the subpockets and pockets. Figure 6.1 shows the highest scoring pocket and sub pocket for Glucocorticoid receptor (GCR). In this example, the subpocket still extends beyond the ligand binding site. As the pocket is considered as a whole, fine details, such as the location of the hotspot, are lost.

Fragment Hotspot Maps provide finer detail than pocket-based methods, as well as a continuum of scores from the hotspot to the full pocket. It is possible to define a "fragment-sized", "lead-sized" or "drug-sized" volume, and select the highest scoring regions occupied by this volume. This gives a better reflection of reality, as ligands can often occupy only part of the pocket, exemplified in figure 6.1. The top left image shows the Fragment Hotspot Map output at a low score contour, highlighting all of the pockets that have been sampled. The top middle image shows the top 300$\mathring{A}^3$, and although the majority of the propensity surrounds the ligand, in this early implementation some propensity can be found in other pockets. This will be prevented in future by only allowing a single volume to contribute to the total volume. The image on the top right shows the highest scoring 150 $\mathring{A}^3$, which is now smaller than the displayed ligand.

Future work will look at the distribution of scores from maps contoured at given volume, exemplified in figure 6.2 (histograms created using a Hotspot API function written by Pete Curran). The distributions will be calculated for sets of tractable and intractable targets, training a support vector machine (SVM) model to classify pockets. This is a similar approach to the one used in VolSite by Desaphy and colleagues [205], however VolSite uses more simple interaction grids and suffers from the same pocket prediction problem identified above. This work should provide a better method for defining a binding site to be used with existing structure-based tractability methods, and offers a potential method for using the maps themselves.

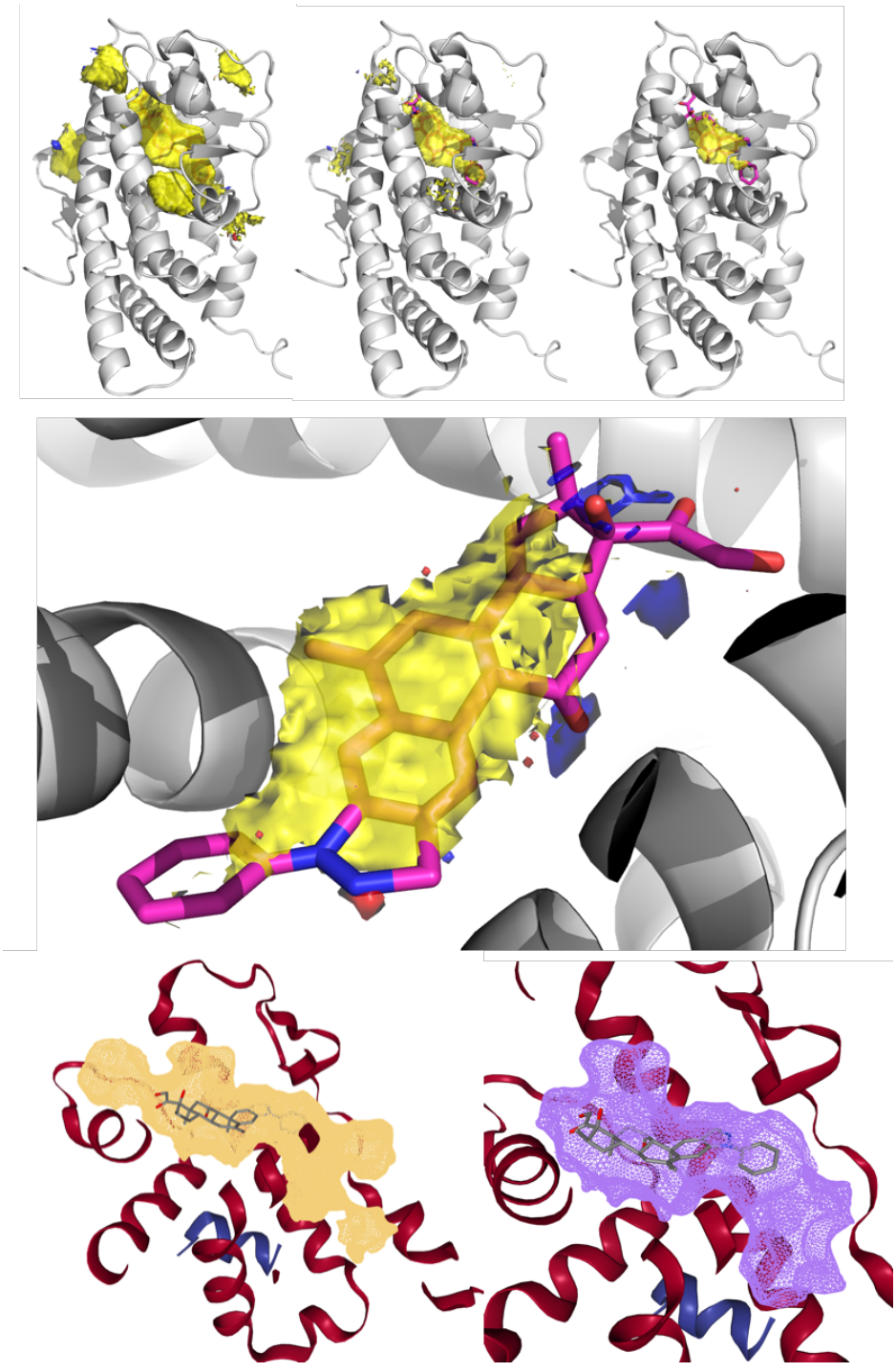


Fig. 6.1 Fragment Hotspot Maps contoured by volume, compared to pocket detection by DoGSiteScorer. (Top) GCR with maps contoured to show all pockets(left) 300 $Å^3$ (middle) and 150 $Å^3$. (Middle) A closer look at the binding site with maps contoured at 150 $Å^3$, ligand shown in magenta sticks. (Bottom) The pocket (orange) and subpocket (purple) for the same structure, as predicted by DoGSiteScorer.

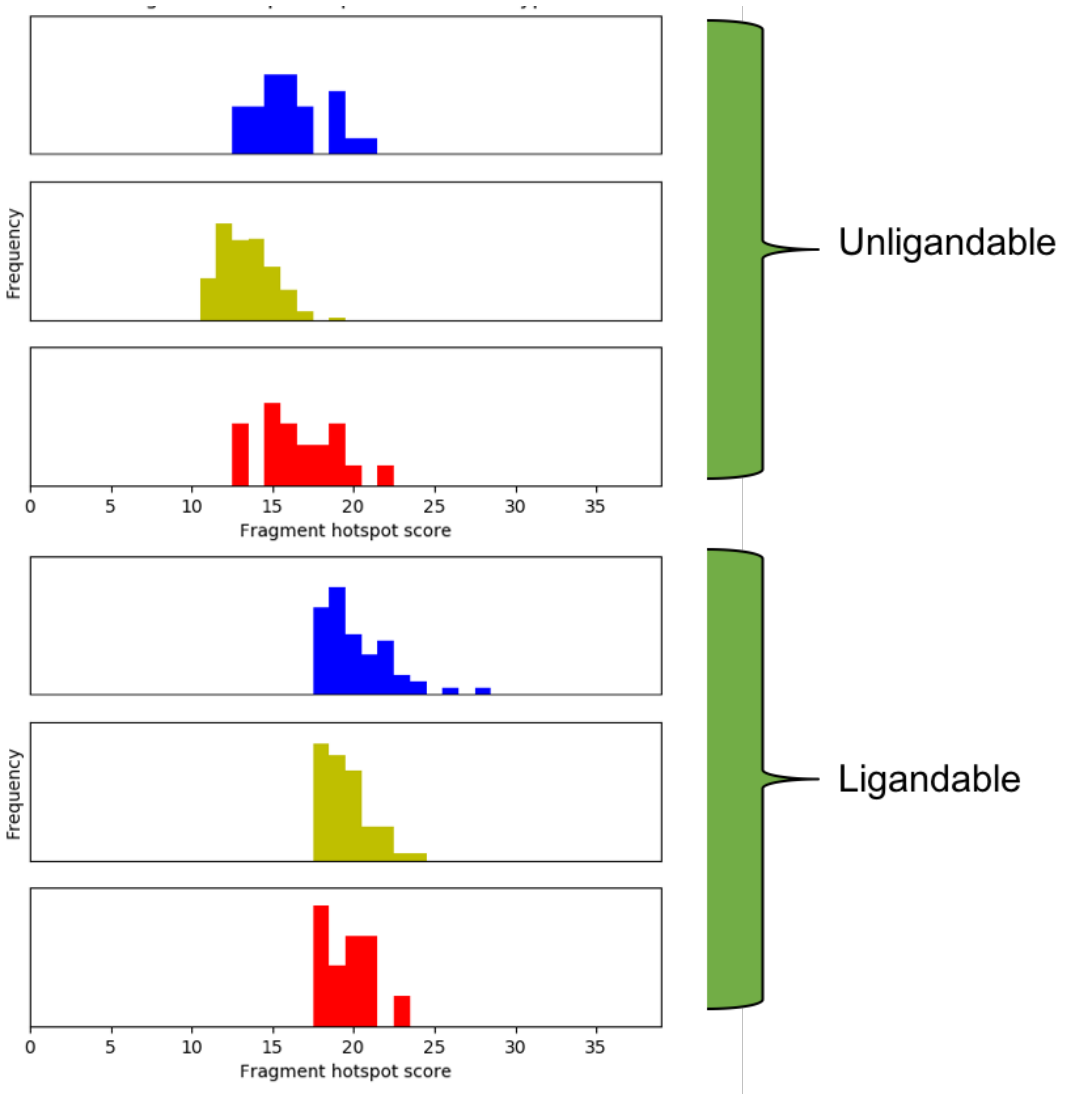


Fig. 6.2 Map score distribution at volume cut-off. Histograms of the top scoring 150 $Å^3$ are shown from two pockets, previously identified as ligandable (GCR, PDB 3BQD) and unligandable (BAZ2B, PDB 3Q2F).

6.3 Decorating Proteomes with Hotspots

Due to the speed of the calculation, it has been possible to calculate Fragment Hotspot Maps for over 9000 ligand-bound structures in the PDB. Although analysis of these data cannot be completed within this PhD project, it will be taken over by another group member. The aim is to refine the cut-off scores for fragments and lead-like molecules defined above by looking at the distribution of scores for both molecule types, and begin to look in to how well the highest scoring interactions match up with chemistry found in the PDB.

As this initial work shows, the ability to automate and the speed of this method allows for analysis of large numbers of structures. Recent work by Somody and colleagues [206] has shown that structural coverage of the human proteome at 30% sequence identity and above is approaching 70%. This corresponds to estimates from the Blundell group that reasonable models are available for approximately 70% of both the human and *Mycobacterium Tuberculosis* proteomes [207]. Although work is required to assess how sensitive Fragment Hotspot Maps are to homologous structures or homology models, it should be possible to decorate a large portion of the human proteome with hotspots. As many of the methods described in chapter 4 require a negligible amount of computation time as compared to the generation of the maps, it will also be possible to run these for each of targets. This includes the generation of pharmacophores, scoring of any ligands and assignment of scores to protein atoms. The work from this project has led to a follow up PhD project titled “Global Analysis of Pharmacophoric Space”. Pharmacophores derived from the hotspot interactions will be used to design and synthesise a chemical library that could deliver hits to any given folded protein target.

Decoration of a proteome also offers the potential to bridge structural and sequence data. As scores can be assigned from the maps to protein atoms, it is possible to annotate the protein sequence with hotspot scores. Residues at functionally important binding sites tend to have greater evolutionary conservation [208], and previous methods have used conservation scores in combination with 3D pocket detection methods to predict ligand binding sites [209]. As an alternative approach, a database of "hotspot motifs" could be created from high scoring regions of annotated sequences. These motifs may have large gaps as distant parts of the sequence are brought close together by the tertiary structure, however these regions are more likely to remain conserved. A query sequence can be compared to the database of hotspot motifs using software such as InterProScan [210]. This method provides two opportunities. Firstly, it can identify targets that may contain a hotspot from their sequence alone. Secondly,

as the hotspot motif will have a link to its original structure, this approach could identify targets with a high homology at the site of the pocket, even when overall sequence identity is low. This work will be explored further in my large scale tractability project at the EBI, where a tractability assessment for targets without a structure is required.

6.4 Decorating MD Trajectories with Hotspots

In addition to looking at larger numbers of crystal structures, this can also be applied to frames extracted from MD simulations. As a proof of concept, 6000 frames from an MD simulation were processed within 48 hours on a Linux workstation. It is possible to visualise the maps for each frame to see how they are affected by the dynamics of the protein. Another way to view the data across the MD trajectory is through the calculation of summary maps. Average maps, shown in figure 6.3, give the average score for each grid point across all the frames. This has the effect of smoothing out the noise in data, removing any artefacts caused by flexible regions being fixed or restrained in the crystal structure. The second is the "maximum" summary map. This captures the highest score achieved by each grid point across the frames. The purpose for this map is to capture hotspots when they appear during the simulation. This work will be continued further by collaborators at UCB.

6.5 Prioritising Fragment Hits

Another use is to prioritise crystallographic fragment screening hits. It is clear that fragments can bind to a protein, without being bound to a hotspot (Alicia Higueruelo, Sherine Thomas and Tom L Blundell, Unpublished). In this situation, the fragment may no longer bind in the same pose or at all upon elaboration. The recently developed PanDDA approach [90] is able to identify a far greater number of fragment binding sites. This method does an excellent job of extracting as much data from the electron density as possible. This gives a good representation of the crystallographic environment, however not all of these sites would be suitable starting points for a drug discovery project. An example is given in figure 6.4, where several fragment-bound structures have been overlaid with the results from a Fragment Hotspot Maps calculation.

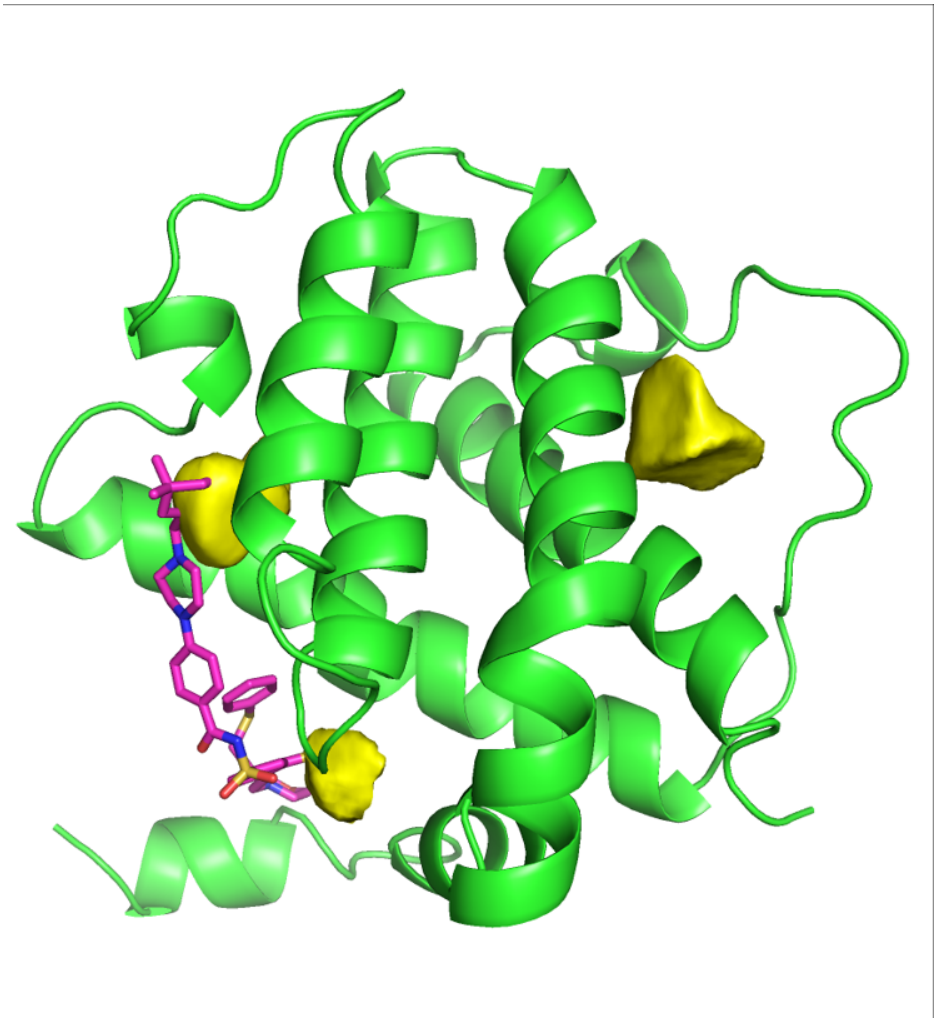


Fig. 6.3 Bcl-xl with the average hydrophobic map from 6000 MD frames. Apo Bcl-xl displayed as a green cartoon with a magenta ligand for reference. The average hydrophobic map is displayed as a yellow surface

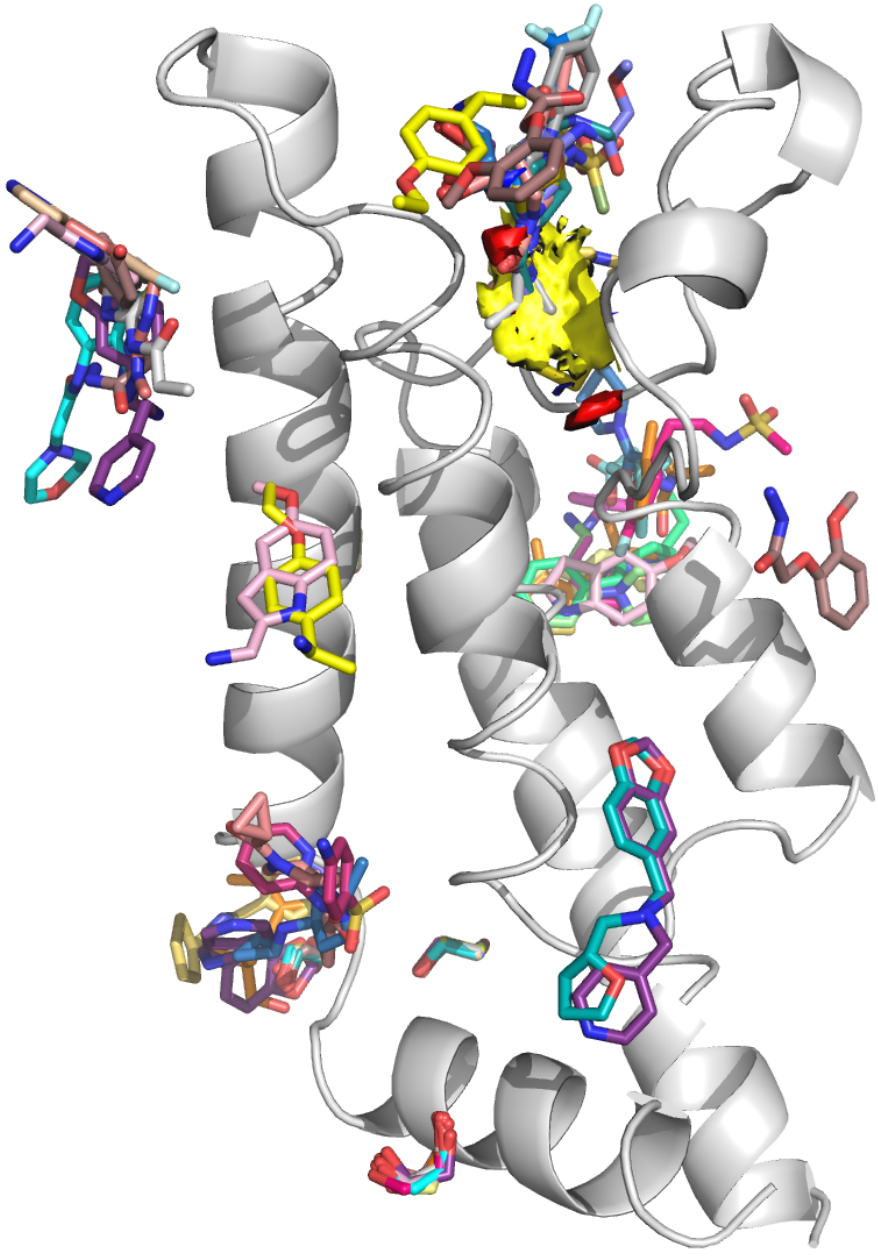


Fig. 6.4 ATAD2 overlaid with fragments identified by PanDDA. Fragment Hotspot Maps are shown contoured at 17. Several fragment binding sites are identified, but many are highly exposed, and not predicted to bind to hotspots.

6.6 Hit-to-lead Development

Fragment Hotspot Maps will be useful to both medicinal chemists and molecular modellers. For the medicinal chemist, the maps will be a visual guide to the most important interactions within the pocket, as demonstrated retrospectively with the examples of protein kinase B and pantothenate synthetase. Once a structure containing a hit is available, it will be easy to determine whether any of the existing groups are suboptimal. The maps will suggest the direction in which the hit should be grown and which types of interaction are required.

An example of Fragment Hotspot Maps guiding fragment growth has been given in figure 6.5. A PyMOL session containing experimental data and the Fragment Hotspot Maps was kindly provided by Sherine Thomas, a fellow member of the Blundell group. I would also like to acknowledge the contributions of Andrew Whitehouse, Alexander Fanourakis, Dr. Anthony Coyne and Prof. Chris Abell at the department of Chemistry to this work. As the compounds are currently unpublished, they have only been represented as wire meshes.

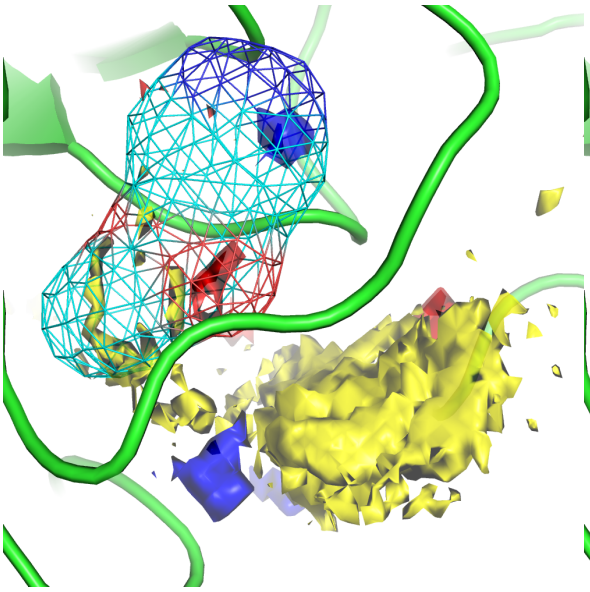
Dissociation constants (K_ds) were calculated for the original fragment hit (compound 1 K_d = 104 μM), first elaboration (compound 2 K_d = 32 μM) and second elaboration (compound 3 K_d = 9 μM). The pocket has four polar interactions with scores greater than 17, two of which are made by the fragment. Compound 2 extends further into the hydrophobic propensity, but does not make any additional polar interactions. Compound 3 extends fully into the identified hydrophobic region, but again does not make any direct polar interactions with the protein. Instead, it appears as though the region of donor propensity, shown at the bottom of each image in figure 6.5, is in fact a water binding site, which bridges an interaction with the ligand. The fact that high scoring interactions sometimes correspond to stable water-binding sites is undesirable behaviour, and will be discussed further in chapter 7. The final polar interaction is beyond the reach of compound 2, and presents an opportunity for further growth. As evidence supporting the hypothesis that this is indeed an important interaction, compound 1 has a second binding mode in the crystal structure where it is found to make this interaction.

It could be argued that you simply need to look at the protein structure to identify where polar interactions are, and growing fragments is a case of picking up these interactions as they are grown into space, however this ignores the varying behaviour of water in the binding site. This example demonstrates two benefits of using Fragment Hotspot Maps. Firstly, it is capable of prioritising the numerous available interactions. Often, as in this case, the highest scoring interactions will be made by the fragment already, however those remaining high scoring interactions offer a chance for improved affinity. Secondly, Fragment Hotspot Maps

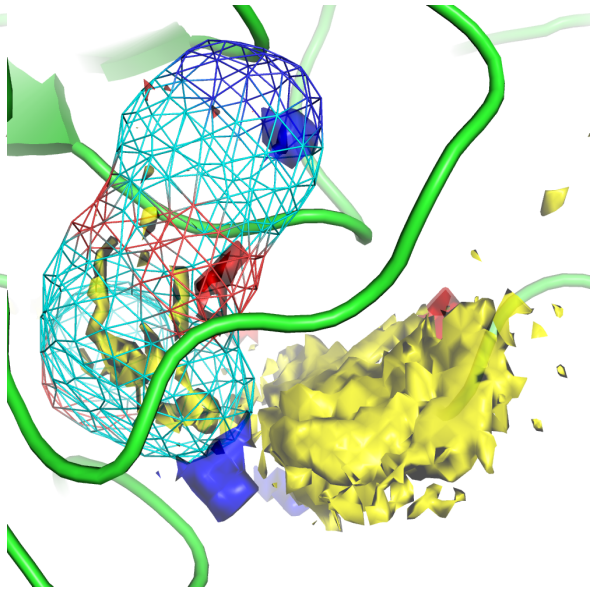
are capable of precisely locating hydrophobic regions which correspond to useful warm spots. In addition to compound 2 in figure 6.5c, this was exemplified retrospectively in chapter 3 looking at the growth of fragments targeting pantothenate synthetase in figure 3.8. In this example, two high scoring hydrophobic regions were identified, and initial work led to growing the fragment into the lower scoring of the two. Later, it was found that growing into the region identified as the highest scoring led to more potent compounds.

6.7 Conclusion

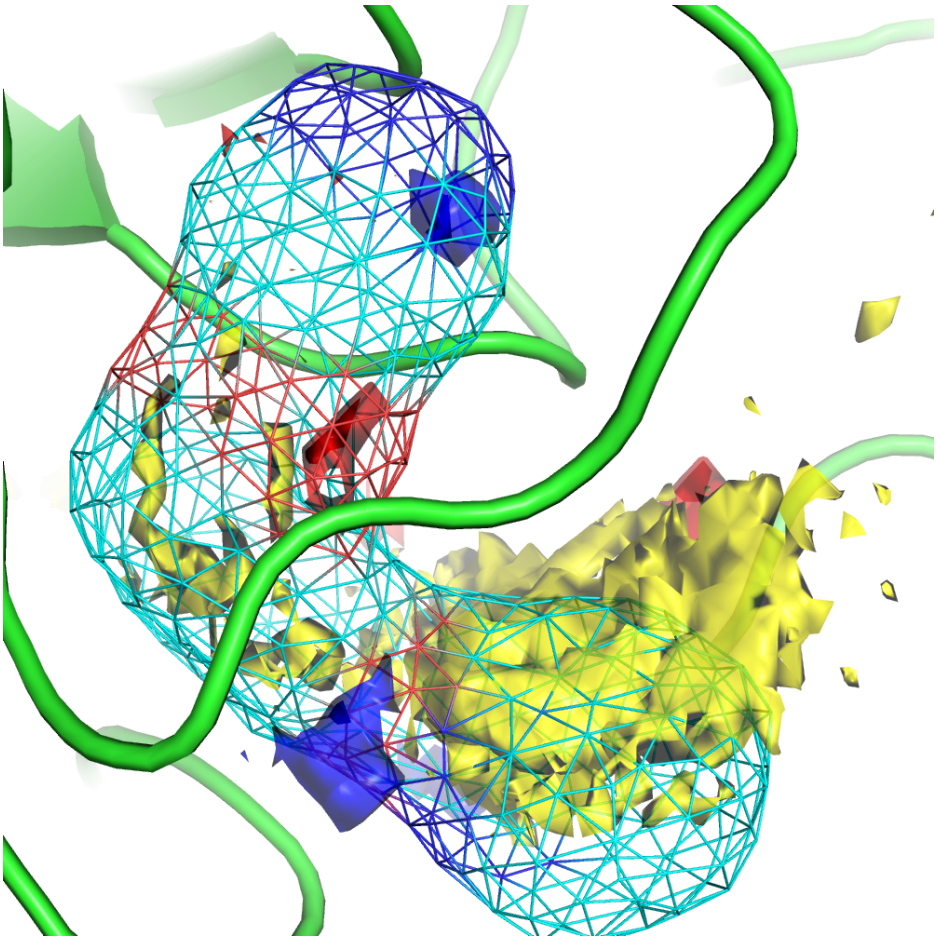
The initial development of the Fragment Hotspot Maps method was completed mid way through this PhD project. As a result, this has given me time to explore how it can be used to help structure-based drug design. This chapter has given a brief overview of ways in which the method is used currently, and how collaborators and I intend to use it in the future. Much of the work moving forward has benefited from the development of the Hotspots API, described in chapter 4. The Hotspots API can be regarded as a tool kit, with functionality ranging from virtual screening to fundamental methods, such as scoring the protein. Having produced this tool kit, other scientists will be able to focus on being creative and working out how they can use the knowledge that Fragment Hotspot Maps provide.



(a) Compound 1: $K_d = 104$ µM



(b) Compound 2: $K_d = 32$ µM



(c) Compound 3: $K_d = 9$ µM

Fig. 6.5 Fragment growing guided by Fragment Hotspot Maps. All maps are contoured at 14, and ligands are shown as a wire mesh. The colour of the mesh represents the atom type, cyan for carbon, blue for nitrogen and red for oxygen.

Chapter 7

Discussion and Conclusions

7.1 Summary

Fragment-based drug design is now a mainstream approach in drug discovery, but few computational methods have been designed specifically to support it. Existing computational hotspot detection methods are either capable of locating fragment binding sites from a global search of a protein [123, 122], or highlighting important interactions within a predefined binding site [143, 130]. Computational methods were able to replicate multiple solvent crystal structure experiments only once solvation was accounted for, explicitly or implicitly [160].

Separate publications looking at unhappy waters [130], fragment-binding sites [132] and hotspots [110] each identified the required protein environment as having hydrophobic enclosure with a mixture of polar and hydrophobic interactions. This prompted the development of the Fragment Hotspot Maps method, which aimed to locate these protein environments. Hotspots were defined as follows:

> **Hotspots** are the minimum binding site that will bind a fragment, maintaining the fragment binding position once it has been elaborated

The calculation of Fragment Hotspot Maps is a CSD-derived computational approach, with a knowledge-based origin that makes it much faster than existing approaches. Calculations can be completed within minutes rather than hours [110]. Fragment Hotspot Maps provide a continuum of scores, which describe the relative likelihood of discovering the given

interaction type at each grid point. Increasing the value at which the maps are contoured locates the most attractive regions and interactions within a pocket.

Given the above hotspot definition, a previously published [132] dataset of fragment-lead pairs was an ideal dataset for testing Fragment Hotspot Maps. The ligand bound protein crystal structures were aligned to apo protein crystal structures, which were used for the calculation of Fragment Hotspot Maps. The maps were able to specifically identify fragment binding sites and their interactions over the larger ligand-binding site.

A web application was developed to provide a simple user interface for both submitting calculations and viewing results. Minimal user input is required to download and prepare a protein from the PBD. NGL viewer [173] is used to display the protein and maps, which can be contoured interactively using a slider. Changing the contouring value modifies the description of how likely the site is going to be a hotspot.

In addition to the web server, the Hotspots API was developed to provide easy and scriptable access to Fragment Hotspot Map calculations. The Hotspots API provides methods for using the information provided by the maps, ranging from simple tasks such as scoring ligand atoms, through to higher-level functions used for virtual screening. The purpose of this work was to create a tool kit to allow other researchers answer their own questions more easily.

To highlight how the Hotspots API can be used to help in structure-based drug design, two of the three virtual screening workflows were applied to the DUD-e test set [195]. Docking was performed both with and without a hotspot-derived constraint, and performance either improved or remained the same. Increasing the weighting of the constraint led to further improvement in performance. The second approach was to use a field-based-ligand screener with modified Fragment Hotspot Maps as input. This was compared to the best case scenario of running the field-based ligand screener with a set of overlaid ligands from multiple protein crystal structures. While often unable to match the early enrichment of the ideal scenario, it showed comparable or better performance at locating novel chemistry. Interestingly, the highly hydrophobic and flexible target GCR performed poorly with docking both in this and previous studies [195] but performed very well using the field-based ligand screener with Fragment Hotspot Map input. This may suggest that this work flow is more suitable for targets of this type, or where docking has previously performed poorly.

7.2 Novelty of Work

Literature discussing the nature of hotspots [132, 110, 130] was used early on in this project to guide the development of the Fragment Hotspot Maps method, and its success adds further confidence to this view of hotspots. While previous methods [143, 122] stated that solvent is required to remove false positives, this work agrees with that of Brenke and colleagues [123], who showed that a "cavity" term can be used to give a similar effect.

Most current hotspot detections methods use MD [133], however FTMap [123] uses a static protein structure and is the closest to the Fragment Hotspot Maps method. It scans the surface of a static protein structure with molecular probes, and is capable of locating fragment-binding sites from a global search of the protein. To my knowledge, FTMap is the only other software reported to be able to do this. The Fragment Hotspot Map method has also demonstrated its ability to locate fragment-binding sites, however it is able to do so in approximately 5-30 minutes, rather than 4-24 hours required by FTMap [110]. This has opened up the possibility of performing wide scale analyses of the human proteome, as discussed in the previous chapter.

Fragment Hotspot Maps highlight the specific interactions within the hotspot, making it possible to generate a pharmacophore model for each of these predicted hotspots. Similar work by Yu and co-workers [211] uses the SILCS MD-based method to identify important interactions within a pre-defined binding site, which they used to generate pharmacophores.

The Hotspots API provides a set of tools to incorporate information about the highest scoring interactions into existing structure-based workflows. While docking requires an expert user to yield the best results, the docking work flow used in chapter 5 could be used to set multiple suggestions for docking constraints and the user select a sensible combination. Earlier this year, around the same time the docking work described in chapter 5 was performed, Arcon and colleagues [142] used the results of mixed solvent molecular dynamics to improve docking. In this work the authors modified the scoring function for AutoDock [212] and assessed their ability to correctly predict the crystallographic poses for two targets. The mixed MD method is used to identify favourable probe binding positions, which are then used to bias the scoring function, however the workflow started from a ligand bound structure (with the ligand removed) and with the binding site defined. Although binding sites are typically known before docking calculations are performed, and is therefore sensible for this use, it demonstrates that this method is not capable of detecting hotspots from a global search.

Fragment Hotspot Map calculations are not only faster than existing hotspot detection methods, but also provide the best features of each: identification of fragment binding sites from a global search of the protein plus highlighting key interactions within the binding site. In order to allow other scientists to get the most out of this program, a programmatic tool kit has be produced to allow them to incorporate Fragment Hotspot Map calculations into their existing work flows.

7.3 Remaining challenges

While the work described in this thesis aims to improve upon existing hotspot detection methods, the provision of interaction information leads to extra challenges. As described above, this work found it was possible to use a cavity/buriedness term in place of solvation. While this is capable of locating unfavourable water regions, the current implementation finds that known structural waters coincide with high scoring regions of the map, suggesting they should be displaced by a fragment. However, Ichihara and colleagues [132] found both types of hydration site could be displaced upon fragment binding (figure 7.1).

- Overall positive excess free energy compared to bulk water (i.e. unhappy water), with positive excess enthalpy (ΔH) and entropy ($-T\Delta S$) contribution. These water molecules are unable to satisfy their hydrogen bonding capabilities as well as in bulk solvent. Any hydrogen bonds made to the protein or neighbouring water molecules restrict the motion of the molecule. Displacement of this water is relatively easy.

- Overall negative excess free energy compared to bulk water (i.e. happy water), with large negative excess enthalpy (ΔH) and large entropy ($-T\Delta S$) contribution. The water molecule makes a very strong interaction with the protein, and its hydrogen bonding capabilities are well satisfied. Due to enthalpy-entropy compensation [180], the water molecule has a large positive excess entropy ($-T\Delta S$) and is held tightly in place. Displacement of this water requires the ligand to precisely replace the interactions of the water molecule.

The fact that waters in energetically favourable positions (happy waters) can be displaced to improve ligand binding may seem counter-intuitive. To rationalise this, the energy of the protein ligand binding event must be considered as a whole. These waters are in a favourable position compared to bulk water due to large but opposing enthalpic and entropic

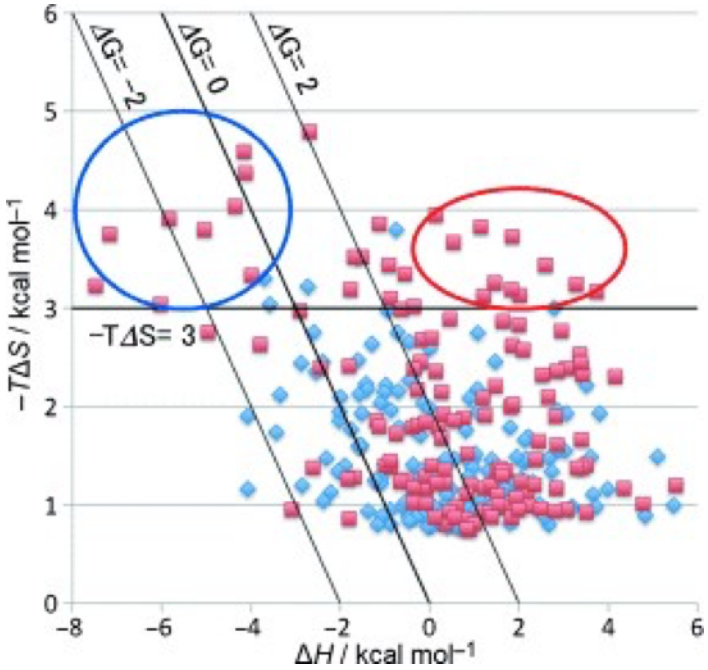


Fig. 7.1 The excess enthalpy (ΔH) and entropy ($-T\Delta S$) of the hydration sites displaced by fragments and lead compound. Hydration sites displaced by fragments are shown in red, lead compounds in blue. The two circled regions represent types of hydration sites exclusively displaced by fragments. Image taken from the original publication [132]

terms. A ligand can benefit from this by replacing the interaction of the water, negating the enthalpic penalty whilst benefiting entropically through the release of the highly constrained water. To do so, ligands must provide precisely matched hydrogen bonds to replace those of the water. On average, fragments make two well-made hydrogen bonds [213], and their simplicity allows them to make the highly geometrically constrained interactions. Replacing the interactions with a ligand negates the lost excess enthalpy upon displacing the water, but benefits entropically from the release of the highly constrained water. Ichihara *et al* [132] state that if a fragment is able to displace such a water, it should be prioritised, as it is easier to grow the fragment into the unhappy water sites, which can often by displaced by hydrophobic groups.

These sites are currently predicted as highly scoring by the Fragment Hotspot Map calculations, which is good for prioritising crystallographic fragment hits, but bad when highlighting interactions that should be targeted by prospective ligands. While some fragments can make these interactions by chance, designing chemistry to match them is far more difficult. These sites are often highly scoring due to the fact that multiple protein hydrogen bonds are directed towards the same region. SuperStar only accounts for the position of the probe atom, and does not account for the direction or quality of a hydrogen bond. As a result, despite using relatively large probes, polar interactions that would be difficult to make in reality are sampled easily and score highly. A potential approach to avoid finding geometrically constrained polar interactions is to post-process the molecular probe poses to eliminate those with poor quality hydrogen bonds. This should have only a small impact on performance, whilst removing the harder-to-reach interactions.

7.4 Concluding Remarks

The field of computational hotspot detection seems to be moving towards MD, which is unsurprising given the recent improvements in computational performance [214], and the importance of protein flexibility in SBDD. Kozakov and colleagues have shown previously that FTMap can correctly identify hotspots even when large conformational changes take place [144]. This does not extend to cryptic sites [215] - the pocket needs to exist - however their method is robust enough to deal with changes in pocket shape. This is also true for Fragment Hotspot Maps; apo structures were used in chapter 3 to identify fragment-binding sites. While hotspots are not sensitive to these changes, the more precise positions of

individual interactions will move with the groups that cause them. This a problem in SBDD as a whole, and existing methods have been developed to deal with this: pharmacophore models allow a search tolerance, docking can use an ensemble of structures or allow side chain flexibility. For now, I feel MD-based hotspot detection methods have not demonstrated the ability to selectively highlight fragment binding sites as effectively as either Fragment Hotspot Maps or FTMap.

The Fragment Hotspot Map method has been developed with FBDD in mind, but is useful for SBDD in general. This can be viewed as a computational equivalent to "fragment-assisted" drug discovery, where knowledge from fragments can be incorporated into molecules discovered by other means [216]. Fragments bound at hotspots are highlighting highly important interactions, which should be optimally matched in larger molecules. Knowledge of these interactions *a priori* can lead to more useful computational experiments ahead of a focussed screen, or to even predict the binding mode of fragments that lack a crystal structure.

The recently published improvement to X-ray crystallography, PanDDA [90], has shown that fragment binding alone is no longer enough to suggest the presence of a hotspot. While PanDDA is very good at using electron density data to show what is present in the environment of the crystal, it presents a challenge: how do you choose which fragments to progress? The computational assessment of fragment binding sites continues to provide important information at multiple stages of the drug discovery process. Fragment Hotspot Maps, in combination with the Hotspots API, can help scientists select tractable pockets, select compounds to screen, prioritise experimental hits and guide hit-to-lead development. With the ever-increasing structural coverage of the human proteome, large-scale structure-based assessment of protein targets is on the horizon. I believe the work described in this thesis can aid not only in this work, but also in the discovery of drugs for the novel targets found.

References

[1] Chris J. Radoux, Tjelvar S. G. Olsson, Will R. Pitt, Colin R. Groom, and Tom L. Blundell. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.*, 59:4314–4325, 2016.

[2] Amy C. Anderson. The process of structure-based drug design. *Chem. Biol.*, 10:787–797, 2003.

[3] Joseph G. Lombardino and John A. Lowe. The role of the medicinal chemist in drug discovery–then and now. *Nat Rev Drug Discov*, 3:853–862, 2004.

[4] B. J. Katz, E. I. Dittmar, and G. E. Ehret. A geochemical review of carbonate source rocks in Italy. *J. Pet. Geol.*, 23:399–424, 2000.

[5] H R Howard, John A. Lowe, T F Seeger, P a Seymour, S H Zorn, P R Maloney, F E Ewing, M E Newman, a W Schmidt, J S Furman, G L Robinson, E Jackson, Christopher N. Johnson, and J Morrone. 3-Benzisothiazolylpiperazine derivatives as potential atypical antipsychotic agents. *J. Med. Chem.*, 39:143–148, 1996.

[6] Richard A. Glennon, R M Slusher, R A Lyon, M Titeler, and J D McKenney. 5-HT1 and 5-HT2 binding characteristics of some quipazine analogues. *J. Med. Chem.*, 29:2375–80, 1986.

[7] H P Rang. The receptor concept: pharmacology's big idea. *Br. J. Pharmacol.*, 147 Suppl:S9–16, 2006.

[8] David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite, and Menelas N Pangalos. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.*, 13:419–31, 2014.

[9] Robert M Plenge. Disciplined approach to drug discovery and early development. *Sci. Transl. Med.*, 8:349ps15, 2016.

[10] M.A. Lindsay. Target discovery. *Nat. Rev. Drug Discov.*, 2:831–838, 2003.

[11] Yongliang Yang, S. James Adelstein, and Amin I. Kassis. Target discovery from data mining approaches. *Drug Discov. Today*, 14:147–154, 2009.

[12] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam MacIejewski, David Arndt, Michael Wilson, Vanessa Neveu, Alexandra Tang, Geraldine Gabriel, Carol Ly, Sakina Adamjee, Zerihun T. Dame, Beomsoo Han, You

Zhou, and David S. Wishart. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.*, 42:D1091–7, 2014.

[13] Hong Yang, Chu Qin, Ying Hong Li, Lin Tao, Jin Zhou, Chun Yan Yu, Feng Xu, Zhe Chen, Feng Zhu, and Yu Zong Chen. Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.*, 44:D1069–D1074, 2016.

[14] Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H. Blicher, Christian Von Mering, Lars J. Jensen, and Peer Bork. STITCH 4: Integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, 42:D401–7, 2014.

[15] M Whirl-Carrillo, E M McDonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, Russ B. Altman, and T E Klein. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, 92:414–7, 2012.

[16] Nikolai Hecker, Jessica Ahmed, Joachim Von Eichborn, Mathias Dunkel, Karel Macha, Andreas Eckert, Michael K. Gilson, Philip E. Bourne, and Robert Preissner. SuperTarget goes quantitative: Update on drug-target interactions. *Nucleic Acids Res.*, 40:D1113–7, 2012.

[17] Gautier Koscielny, Peter An, Denise Carvalho-Silva, Jennifer A Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, Nikiforos Karamanis, Michael Maguire, Eliseo Papa, Andrea Pierleoni, Miguel Pignatelli, Theo Platt, Francis Rowland, Priyanka Wankar, A. Patrícia Bento, Tony Burdett, Antonio Fabregat, Simon Forbes, Anna Gaulton, Cristina Yenyxe Gonzalez, Henning Hermjakob, Anne Hersey, Steven Jupe, Şenay Kafkas, Maria Keays, Catherine Leroy, Francisco-Javier Lopez, Maria Paula Magarinos, James Malone, Johanna McEntyre, Alfonso Munoz-Pomer Fuentes, Claire O'Donovan, Irene Papatheodorou, Helen Parkinson, Barbara Palka, Justin Paschall, Robert Petryszak, Naruemon Pratanwanich, Sirarat Sarntivijal, Gary Saunders, Konstantinos Sidiropoulos, Thomas Smith, Zbyslaw Sondka, Oliver Stegle, Y Amy Tang, Edward Turner, Brendan Vaughan, Olga Vrousgou, Xavier Watkins, Maria-Jesus Martin, Philippe Sanseau, Jessica Vamathevan, Ewan Birney, Jeffrey Barrett, and Ian Dunham. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, 45:gkw1055, 2016.

[18] Prisca Honore, Karen Kage, Joseph Mikusa, Andrew T. Watt, Joseph F. Johnston, Jacqueline R. Wyatt, Connie R. Faltynek, Michael F. Jarvis, and Kevin Lynch. Analgesic profile of intrathecal P2X3 antisense oligonucleotide treatment in chronic inflammatory and neuropathic pain states in rats. *Pain*, 99:11–19, 2002.

[19] Caitlin Smith. Drug target validation: Hitting the target. *Nature*, 422:341, 343, 345 passim, 2003.

[20] Aroon D. Hingorani, Valerie Kuan, Chris Finan, Felix A. Krüger, Anna Gaulton, Sandesh Chopade, Reecha Sofat, Raymond J Macallister, John P. Overington, Harry Hemingway, Spiros Denaxas, David Prieto, and Juan Pablo Casas. Flipping the odds of drug development success through human genomics. *bioRxiv*, 2017.

[21] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 64:4–17, 2012.

[22] Andrew L. Hopkins and Colin R. Groom. The druggable genome. *Nat. Rev. Drug Discov.*, 1:727–30, 2002.

[23] Chris Finan, Anna Gaulton, Felix A. Krüger, Tom Lumbers, Tina Shah, Jorgen Engmann, Luana Galver, Ryan Kelly, Anneli Karlsson, Rita Santos, John P. Overington, Aroon D. Hingorani, Juan Pablo Casas, R. Thomas Lumbers, and Ryan Kelley. The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.*, 9:066027, 2017.

[24] Fredrik N B Edfeldt, Rutger H A Folmer, and Alexander L. Breeze. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov. Today*, 16:284–287, 2011.

[25] Philip J. Hajduk, Jeffrey R Huth, and Stephen W. Fesik. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, 48:2518–25, 2005.

[26] Alan C Cheng, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.*, 25:71–5, 2007.

[27] Udo Bauer and Alexander L. Breeze. "Ligandability" of Drug Targets: Assessment of Chemical Tractability via Experimental and In Silico Approaches. In *Lead Gener. Methods, Strateg. Case Stud.*, pages 35–62. Wiley-VCH Verlag GmbH & Co. KGaA, 2016.

[28] James A Wells and Christopher L McClendon. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450:1001–9, 2007.

[29] Andrea G. Cochran. Antagonists of protein-protein interactions. *Chem. Biol.*, 7:R85–R94, 2000.

[30] Robin W. Spencer. High-throughput screening of historic collections: Observations on file size, biological targets, and file diversity. *Biotechnol. Bioeng.*, 61:61–67, 1998.

[31] J. Janin and C. Chothia. The structure of protein-protein recognition sites. *J. Biol. Chem.*, 265:16027–16030, 1990.

[32] Warren L DeLano. Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.*, 12:14–20, 2002.

[33] Duncan E. Scott, Matthias T. Ehebauer, Tara Pukala, May Marsh, Tom L. Blundell, Ashok R. Venkitaraman, Chris Abell, and Marko Hyvönen. Using a Fragment-Based Approach To Target Protein-Protein Interactions. *ChemBioChem*, 14:332–342, 2013.

[34] Peter Schmidtke and Xavier Barril. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, 53:5858–67, 2010.

[35] H. M. Berman. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[36] Thomas A Halgren. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, 49:377–89, 2009.

[37] Andrea Volkamer, Daniel Kuhn, Thomas Grombacher, Friedrich Rippmann, and Matthias Rarey. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.*, 52:360–72, 2012.

[38] Johann Gasteiger. The Scope of Chemoinformatics. In *Handb. Chemoinformatics*, volume 1, pages 1–5. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008.

[39] David Weininger. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[40] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. InChI - The worldwide chemical structure identifier standard. *J. Cheminform.*, 5:7, 2013.

[41] David Weininger, Arthur Weininger, and Joseph L. Weininger. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[42] Jonathan S. Lindsey. A retrospective on the automation of laboratory synthetic chemistry. *Chemom. Intell. Lab. Syst.*, 17:15–45, 1992.

[43] Ronald Frank. Simultaneous and combinatorial chemical synthesis techniques for the generation and screening of molecular diversity. *J. Biotechnol.*, 41:259–272, 1995.

[44] J F Cargill and M Lebl. New methods in combinatorial chemistry-robotics and parallel synthesis. *Curr. Opin. Chem. Biol.*, 1:67–71, 1997.

[45] John Major. Challenges and Opportunities in High Throughput Screening: Implications for New Technologies. *J. Biomol. Screen.*, 3:13–17, 1998.

[46] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, Ian Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, Paul L Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer,

J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H M Yang, J Yu, J Wang, G Y Huang, J Gu, L Hood, L Rowen, A Madan, S Z Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, Roland Schulz, B A Roe, F Chen, H Q Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blocker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, Alex Bateman, S Batzoglou, Ewan Birney, Peer Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G R Gilbert, C Harmon, Y Hayashizaki, D Haussler, Henning Hermjakob, K Hokamp, W H Jang, L S Johnson, T. A. Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J R Schultz, G Slater, August B Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, and M J Morgan. Initial sequencing and analysis of the human genome. *Nat.*, 409:860–921, 2001.

[47] Jonathan J. Burbaum. Miniaturization technologies in HTS: How fast, how small, how soon? *Drug Discov. Today*, 3:313–322, 1998.

[48] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren V S Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, and Ulrich Schopfer. Impact of high-throughput screening. *Nature*, 10:188–195, 2011.

[49] Mike M. Hann and Tudor I. Oprea. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.*, 8:255–263, 2004.

[50] Tobias Fink and Jean Louis Raymond. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discove. *J. Chem. Inf. Model.*, 47:342–353, 2007.

[51] Stephen J. Lane, Drake S. Eggleston, Keith A. Brinded, John C. Hollerton, Nicholas L. Taylor, and Simon A. Readshaw. Defining and maintaining a high quality screening collection: The GSK experience. *Drug Discov. Today*, 11:267–272, 2006.

[52] Lu Tan, Eugen Lounkine, and Jürgen Bajorath. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.*, 48:2308–2312, 2008.

[53] C John Harris, Richard D Hill, David W Sheppard, Martin J Slater, and Pieter F W Stouten. The design and application of target-focused compound libraries. *Comb. Chem. High Throughput Screen.*, 14:521–31, 2011.

[54] T. A. Jones. A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.*, 11:268–272, 1978.

[55] Tom L. Blundell, B. L. Sibanda, and L Pearl. Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature*, 304:273–5, 1983.

[56] Tom L. Blundell. Protein crystallography and drug discovery: Recollections of knowledge exchange between academia and industry. *IUCrJ*, 4:308–321, 2017.

[57] M. J. Sutcliffe, F. R F Hayes, and Tom L. Blundell. Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. *Protein Eng. Des. Sel.*, 1:385–392, 1987.

[58] M. J. Sutcliffe, I. Haneef, D. Carney, and Tom L. Blundell. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng. Des. Sel.*, 1:377–384, 1987.

[59] Andrej Šali and Tom L. Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.*, 234:779–815, 1993.

[60] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[61] Paul D. Leeson and Brian Springthorpe. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.*, 6:881–890, 2007.

[62] György M. Keseru and Gergely M Makara. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.*, 8:203–212, 2009.

[63] W. Patrick Walters, Jeremy Green, Jonathan R. Weiss, and Mark A. Murcko. What do medicinal chemists actually make? A 50-year retrospective. *J. Med. Chem.*, 54:6405–6416, 2011.

[64] Miles Congreve, Robin A E Carr, Christopher W. Murray, and Harren Jhoti. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today*, 8:876–877, 2003.

[65] David C. Rees, Miles Congreve, Christopher W. Murray, and Robin A E Carr. Fragment-based lead discovery. *Nat. Rev. Drug Discov.*, 3:660–72, 2004.

[66] Mike M. Hann, Andrew R Leach, and G. Harper. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Model.*, 41:856–864, 2001.

[67] Irwin D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proc. Natl. Acad. Sci.*, 96:9997–10002, 1999.

[68] Diane Joseph-McCarthy, Arthur J Campbell, Gunther Kern, and Demetri T. Moustakas. Fragment-based lead discovery and design. *J. Chem. Inf. Model.*, 54:693–704, 2014.

[69] Marcel L. Verdonk and David C. Rees. Group efficiency: a guideline for hits-to-leads chemistry. *ChemMedChem*, 3:1179–80, 2008.

[70] Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51:4990–5003, 2012.

[71] H. Leonardo Silvestre, Tom L. Blundell, Chris Abell, and Alessio Ciulli. Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proc. Natl. Acad. Sci.*, 110:12984–12989, 2013.

[72] Ellene H Mashalidis, Paweł Śledź, Steffen Lang, and Chris Abell. A three-stage biophysical screening cascade for fragment-based drug discovery. *Nat. Protoc.*, 8:2309–2324, 2013.

[73] John F. Brandts and Lung Nan Lin. Study of strong to ultratight protein interactions using differential scanning calorimetry. *Biochemistry*, 29:6927–6940, 1990.

[74] Geoffrey A Holdgate and Walter H J Ward. Measurements of binding thermodynamics in drug discovery. *Drug Discov. Today*, 10:1543–50, 2005.

[75] Sau-Mei Leung, Guillermo Senisterra, Kenneth P. Ritchie, Seth E. Sadis, James R. Lepock, and Lawrence E. Hightower. Thermal activation of the bovine Hsc70 molecular chaperone at physiological temperatures: physical evidence of a molecular thermometer. *Cell Stress Chaperones*, 1:78, 1996.

[76] Jennifer Hyde, Andrew C Braisted, Mike Randal, and Michelle R Arkin. Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2. *Biochemistry*, 42:6475–83, 2003.

[77] Brian C Raimundo, Johan D Oslob, Andrew C Braisted, Jennifer Hyde, Robert S McDowell, Mike Randal, Nathan D Waal, Jennifer Wilkinson, Chul H Yu, and Michelle R Arkin. Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: an approach for inhibiting protein-protein interactions. *J. Med. Chem.*, 47:3111–30, 2004.

[78] S. B. Shuker, Philip J. Hajduk, R. P. Meadows, and Stephen W. Fesik. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science (80-.).*, 274:1531–1534, 1996.

[79] Tom L. Blundell, Harren Jhoti, and Chris Abell. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.*, 1:45–54, 2002.

[80] Christopher W. Murray, Maria G Carr, Owen Callaghan, Gianni Chessari, Miles Congreve, Suzanna Cowan, Joseph E Coyle, Robert Downham, Eva Figueroa, Martyn Frederickson, Brent Graham, Rachel L McMenamin, Michael A O'Brien, Sahil Patel, Theresa R Phillips, Glyn Williams, Andrew J Woodhead, and Alison J-A Woolford. Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. *J. Med. Chem.*, 53:5942–55, 2010.

[81] Harren Jhoti. High-throughput structural proteomics using x-rays. *Trends Biotechnol.*, 19:S67–S71, 2001.

[82] V L Nienaber, Paul L Richardson, V Klighofer, J J Bouska, V L Giranda, and J Greer. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.*, 18:1105–1108, 2000.

[83] Disha Patel, Joseph D. Bauman, and Eddy Arnold. Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery. *Prog. Biophys. Mol. Biol.*, 116:92–100, 2014.

[84] Michael J. Hartshorn, Christopher W. Murray, Anne Cleasby, Martyn Frederickson, Ian J. Tickle, and Harren Jhoti. Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.*, 48:403–13, 2005.

[85] Johannes Schiebel, Nedyalka Radeva, Stefan G. Krimmer, Xiaojie Wang, Martin Stieler, Frederik R. Ehrmann, Kan Fu, Alexander Metz, Franziska U. Huschmann, Manfred S. Weiss, Uwe Mueller, Andreas Heine, and Gerhard Klebe. Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study. *ACS Chem. Biol.*, 11:1693–1701, 2016.

[86] Helene Köster, Tobias Craan, Sascha Brass, Christian Herhaus, Matthias Zentgraf, Lars Neumann, Andreas Heine, and Gerhard Klebe. A small nonrule of 3 compatible fragment library provides high hit rate of endothiapepsin crystal structures with various fragment chemotypes. *J. Med. Chem.*, 54:7784–96, 2011.

[87] Johannes Schiebel, Nedyalka Radeva, Helene Köster, Alexander Metz, Timo Krotzky, Maren Kuhnert, Wibke E. Diederich, Andreas Heine, Lars Neumann, Cedric Atmanene, Dominique Roecklin, Valérie Vivat-Hannah, Jean Paul Renaud, Robert Meinecke, Nina Schlinck, Astrid Sitte, Franziska Popp, Markus Zeeb, and Gerhard Klebe. One Question, Multiple Answers: Biochemical and Biophysical Screening Methods Retrieve Deviating Fragment Hit Lists. *ChemMedChem*, 10:1511–1521, 2015.

[88] Nedyalka Radeva, Stefan G. Krimmer, Martin Stieler, Kan Fu, Xiaojie Wang, Frederik R. Ehrmann, Alexander Metz, Franziska U. Huschmann, Manfred S. Weiss, Uwe Mueller, Johannes Schiebel, Andreas Heine, and Gerhard Klebe. Experimental active-site mapping by fragments: Hot spots remote from the catalytic center of endothiapepsin. *J. Med. Chem.*, 59:7561–7575, 2016.

[89] Frank von Delft. http://www.diamond.ac.uk/Beamlines/Mx/Fragment-Screening.html, 2017.

[90] Nicholas M. Pearce, Anthony R. Bradley, Tobias Krojer, Brian D. Marsden, Charlotte M. Deane, and Frank Von Delft. Partial-occupancy binders identified by the Pan-Dataset Density Analysis method offer new chemical opportunities and reveal cryptic binding sites. *Struct. Dyn.*, 4:032104, 2017.

[91] Christopher W. Murray and David C. Rees. The rise of fragment-based drug discovery. *Nat. Chem.*, 1:187–92, 2009.

[92] Barbara Becattini, Carsten Culmsee, Marilisa Leone, Dayong Zhai, Xiyun Zhang, Kevin J Crowell, Michele F Rega, Stefan Landshamer, John C Reed, Nikolaus Plesnila, and Maurizio Pellecchia. Structure-activity relationships by interligand NOE-based

design and synthesis of antiapoptotic compounds targeting Bid. *Proc. Natl. Acad. Sci. U. S. A.*, 103:12602–6, 2006.

[93] Georges Lauri and Paul A. Bartlett. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput. Aided. Mol. Des.*, 8:51–66, 1994.

[94] Michael B. Eisen, Don C. Wiley, Martin Karplus, and Roderick E. Hubbard. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins Struct. Funct. Bioinforma.*, 19:199–221, 1994.

[95] David A. Pearlman and Mark A. Murcko. CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.*, 39:1651–1663, 1996.

[96] Philip J. Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.*, 6:211–219, 2007.

[97] Petar O. Nikiforov, Sachin Surade, Michal Blaszczyk, Vincent Delorme, Priscille Brodin, Alain R. Baulard, Tom L. Blundell, and Chris Abell. A fragment merging approach towards the development of small molecule inhibitors of Mycobacterium tuberculosis EthR for use as ethionamide boosters. *Org. Biomol. Chem.*, 14:2318–2326, 2016.

[98] Alessio Ciulli and Chris Abell. Fragment-based approaches to enzyme inhibition. *Curr. Opin. Biotechnol.*, 18:489–96, 2007.

[99] Gordon Saxty, Steven J Woodhead, Valerio Berdini, Thomas G Davies, Marcel L. Verdonk, Paul G. Wyatt, Robert G Boyle, David Barford, Robert Downham, Michelle D Garrett, and Robin A E Carr. Identification of inhibitors of protein kinase B using fragment-based lead discovery. *J. Med. Chem.*, 50:2293–6, 2007.

[100] Karen N. Allen, Cornelia R. Bellamacina, Xiaochun Ding, Constance J. Jeffery, Carla Mattos, Gregory A. Petsko, and Dagmar Ringe. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins †. *J. Phys. Chem.*, 100:2605–2611, 1996.

[101] Andrew Miranker and Martin Karplus. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins Struct. Funct. Bioinforma.*, 11:29–34, 1991.

[102] Paul A Rejto and G M Verkhivker. Unraveling principles of lead discovery: from unfrustrated energy landscapes to novel molecular anchors. *Proc. Natl. Acad. Sci. U. S. A.*, 93:8945–50, 1996.

[103] Andrew C. English, Sarah H. Done, Leo S D Caves, Colin R. Groom, and Roderick E. Hubbard. Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins Struct. Funct. Genet.*, 37:628–640, 1999.

[104] Dagmar Ringe and Carla Mattos. Analysis of the binding surfaces of proteins. *Med. Res. Rev.*, 19:321–331, 1999.

[105] Andrew C. English, Colin R. Groom, and R. E. Hubbard. Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng. Des. Sel.*, 14:47–59, 2001.

[106] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.

[107] Tim Clackson and James A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science (80-.).*, 267:383–386, 1995.

[108] Philip J. Hajduk. DRUG DESIGN: Discovering High-Affinity Ligands for Proteins. *Science (80-.).*, 278:497–499, 1997.

[109] Christopher W. Murray and Marcel L. Verdonk. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided. Mol. Des.*, 16:741–753, 2002.

[110] Dima Kozakov, Laurie E Grove, David R. Hall, Tanggis Bohnuud, Scott E Mottarella, Lingqi Luo, Bing Xia, Dmitri Beglov, and Sandor Vajda. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.*, 10:733–55, 2015.

[111] Olga Kennard. *From private data to public knowledge.* London: Portland Press Ltd, 1997.

[112] Ian J. Bruno, Jason C. Cole, Magnus Kessler, Jie Luo, W. D Sam Momerwell, Lucy H. Purkis, Barry R. Smith, Robin Taylor, Richard I. Cooper, Stephanie E. Harris, and A. Guy Orpen. Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.*, 44:2133–2144, 2004.

[113] Colin R. Groom. Small Molecule Crystal Structures in Drug Discovery. pages 107–114. Springer, Dordrecht, 2015.

[114] Ian J. Bruno, Jason C. Cole, Jos P. M. Lommerse, R. Scott Rowland, Robin Taylor, and Marcel L. Verdonk. IsoStar: a library of information about nonbonded interactions. *J. Comput. Aided. Mol. Des.*, 11:525–537, 1997.

[115] Marcel L. Verdonk, Jason C. Cole, and Robin Taylor. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.*, 289:1093–108, 1999.

[116] Peter A. Wood, Tjelvar S. G. Olsson, Jason C. Cole, Simon J. Cottrell, Neil Feeder, Peter T. A. Galek, Colin R. Groom, and Elna Pidcock. Evaluation of molecular crystal structures using Full Interaction Maps. *CrystEngComm*, 15:65, 2013.

[117] Robin Taylor. The hydrogen bond between N-H or O-H and organic fluorine: Favourable yes, competitive no. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.*, 73:474–488, 2017.

[118] Harry C. Jubb, Alicia P. Higueruelo, Bernardo Ochoa-Montaño, Will R. Pitt, David B. Ascher, and Tom L. Blundell. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.*, 429:365–371, 2017.

[119] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge structural database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.*, 72:171–179, 2016.

[120] Frank H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B Struct. Sci.*, 58:380–388, 2002.

[121] Carla Mattos, Cornelia R. Bellamacina, Ezra Peisach, Antonio Pereira, Dennis Vitkup, Gregory A. Petsko, and Dagmar Ringe. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J. Mol. Biol.*, 357:1471–1482, 2006.

[122] John L. Kulp, David L Pompliano, and Frank Guarnieri. Diverse fragment clustering and water exclusion identify protein hot spots. *J. Am. Chem. Soc.*, 133:10740–3, 2011.

[123] Ryan Brenke, Dima Kozakov, Gwo-Yu Chuang, Dmitri Beglov, David R. Hall, Melissa R. Landon, Carla Mattos, and Sandor Vajda. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*, 25:621–7, 2009.

[124] David R. Hall, Laurie E Grove, Christine Yueh, Chi Ho Ngan, Dima Kozakov, and Sandor Vajda. Robust identification of binding hot spots using continuum electrostatics: application to hen egg-white lysozyme. *J. Am. Chem. Soc.*, 133:20668–71, 2011.

[125] David R. Hall, Chi Ho Ngan, Brandon S Zerbe, Dima Kozakov, and Sandor Vajda. Hot spot analysis for driving the development of hits into leads in fragment-based drug discovery. *J. Chem. Inf. Model.*, 52:199–209, 2012.

[126] Frank Guarnieri and Mihaly Mezei. Simulated annealing of chemical potential: A general procedure for leading bound waters. Application to the study of the differential hydration propensities of the major and minor grooves of DNA. *J. Am. Chem. Soc.*, 118:8493–8494, 1996.

[127] Phillip W. Snyder, Matthew R. Lockett, Demetri T. Moustakas, and George M. Whitesides. Is it the shape of the cavity, or the shape of the water in the cavity? *Eur. Phys. J. Spec. Top.*, 223:853–891, 2014.

[128] Caterina Barillari, Justine Taylor, Russell Viner, and Jonathan W Essex. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.*, 129:2577–87, 2007.

[129] Stephen P. Andrews, Jonathan S. Mason, Edward Hurrell, and Miles Congreve. Structure-based drug design of chromone antagonists of the adenosine A2A receptor. *Medchemcomm*, 5:571, 2014.

[130] Tom Young, Robert Abel, Byungchan Kim, Bruce J Berne, and Richard A. Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U. S. A.*, 104:808–13, 2007.

[131] Y K Cheng and P J Rossky. Surface topography dependence of biomolecular hydrophobic hydration. *Nature*, 392:696–9, 1998.

[132] Osamu Ichihara, Yuzo Shimada, and Daisuke Yoshidome. The importance of hydration thermodynamics in fragment-to-lead optimization. *ChemMedChem*, 9:2708–17, 2014.

[133] Phani Ghanakota and Heather A. Carlson. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.*, 59:10383–10399, 2016.

[134] Jesus Seco, F. Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *J. Med. Chem.*, 52:2363–2371, 2009.

[135] Katrina W. Lexa and Heather A. Carlson. Full protein flexibility is essential for proper hot-spot mapping. *J. Am. Chem. Soc.*, 133:200–202, 2011.

[136] Chao-Yie Yang and Shaomeng Wang. Computational analysis of protein hotspots. *ACS Med. Chem. Lett.*, 1:125–129, 2010.

[137] Ahmet Bakan, Neysa Nevins, Ami S. Lakdawala, and Ivet Bahar. Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. *J. Chem. Theory Comput.*, 8:2435–2447, 2012.

[138] Danzhi Huang and Amedeo Caflisch. Small molecule binding to proteins: Affinity and binding/unbinding dynamics from atomistic simulations. *ChemMedChem*, 6:1578–1580, 2011.

[139] Yaw Sing Tan, Paweł Śledź, Steffen Lang, Christopher J. Stubbs, David R. Spring, Chris Abell, and Robert B. Best. Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. *Angew. Chemie - Int. Ed.*, 51:10078–10081, 2012.

[140] Nicolas Basse, Joel L. Kaar, Giovanni Settanni, Andreas C. Joerger, Trevor J. Rutherford, and Alan R. Fersht. Toward the Rational Design of p53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem. Biol.*, 17:46–56, 2010.

[141] Priyanka Prakash, John F. Hancock, and Alemayehu A. Gorfe. Binding hotspots on K-ras: Consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. *Proteins Struct. Funct. Bioinforma.*, 83:898–909, 2015.

[142] Juan Pablo Arcon, Lucas A. Defelipe, Carlos P. Modenutti, Elias D. López, Daniel Alvarez-Garcia, Xavier Barril, Adrián G. Turjanski, and Marcelo A. Martí. Molecular Dynamics in Mixed Solvents Reveals Protein-Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. *J. Chem. Inf. Model.*, 57:846–863, 2017.

[143] Daniel Alvarez-Garcia and Xavier Barril. Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. *J. Med. Chem.*, 57:8530–9, 2014.

[144] Dima Kozakov, David R. Hall, Gwo-Yu Chuang, R. Cencic, Ryan Brenke, L. E. Grove, Dmitri Beglov, J. Pelletier, Adrian Whitty, and Sandor Vajda. Structural conservation of druggable hot spots in protein-protein interfaces. *Proc. Natl. Acad. Sci.*, 108:13528–13533, 2011.

[145] Daniel K. Treiber and Neil P. Shah. Ins and outs of kinase DFG motifs. *Chem. Biol.*, 20:745–746, 2013.

[146] David R. Hall and Istvan J Enyedy. Computational solvent mapping in structure-based drug design. *Future Med. Chem.*, 7:337–353, 2015.

[147] Rafael C Bernardi, Marcelo C R Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems . *BBA - Gen. Subj.*, 1850:872–877, 2014.

[148] Andrea Cavalli, Andrea Spitaleri, Giorgio Saladino, and Francesco Luigi Gervasio. Investigating drug-target association and dissociation mechanisms using metadynamics-based algorithms. *Acc. Chem. Res.*, 48:277–285, 2015.

[149] Matteo Masetti, Andrea Cavalli, Maurizio Recanatini, and Francesco Luigi Gervasio. Exploring complex protein-ligand recognition mechanisms with coarse metadynamics. *J. Phys. Chem. B*, 113:4807–4816, 2009.

[150] Fabio Pietrucci, Attilio Vittorio Vargiu, and Agata Kranjc. HIV-1 Protease Dimerization Dynamics Reveals a Transient Druggable Binding Pocket at the Interface. *Sci. Rep.*, 5:18555, 2015.

[151] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[152] Andrew G. Leach, Huw D Jones, David A. Cosgrove, Peter W Kenny, Linette Ruston, Philip MacFaul, J Matthew Wood, Nicola Colclough, and Brian Law. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.*, 49:6672–82, 2006.

[153] Jameed Hussain and Ceara Rea. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.*, 50:339–48, 2010.

[154] Shana L Posy, Brian L Claus, Matt E Pokross, and Stephen R Johnson. 3D matched pairs: integrating ligand- and structure-based knowledge for ligand design and receptor annotation. *J. Chem. Inf. Model.*, 53:1576–88, 2013.

[155] Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak. VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization. *J. Med. Chem.*, 56:5203–7, 2013.

[156] Anna Gaulton, Louisa J. Bellis, A. Patrícia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40:D1100–7, 2012.

[157] Stefan Geschwindner, Johan Ulander, and Patrik Johansson. Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip? *J. Med. Chem.*, 58:6321–6335, 2015.

[158] Shipra Malhotra and John Karanicolas. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *J. Med. Chem.*, 60:128–145, 2017.

[159] Sinisa Vukovic, Paul E. Brennan, and David J. Huggins. Exploring the role of water in molecular recognition: predicting protein ligandability using a combinatorial search of surface hydration sites. *J. Phys. Condens. Matter*, 28:344007, 2016.

[160] Sheldon Dennis, Tamas Kortvelyesi, and Sandor Vajda. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 99:4290–5, 2002.

[161] Michael K. Gilson and Barry H. Honig. Energetics of charge-charge interactions in proteins. *Proteins Struct. Funct. Bioinforma.*, 3:32–52, 1988.

[162] Barry H. Honig and Anthony Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–9, 1995.

[163] Manfred Hendlich, Friedrich Rippmann, and G Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, 15:359–63, 389, 1997.

[164] Stefan Bietz, Sascha Urbaczek, Benjamin Schulz, and Matthias Rarey. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminform.*, 6:12, 2014.

[165] S. Bressanelli, L. Tomei, A. Roussel, I. Incitti, R. L. Vitale, M. Mathieu, R. De Francesco, and F. A. Rey. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl. Acad. Sci.*, 96:13034–13039, 1999.

[166] Bernhard Baum, Laveena Muley, Michael Smolinski, Andreas Heine, David Hangauer, and Gerhard Klebe. Non-additivity of Functional Group Contributions in Protein–Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. *J. Mol. Biol.*, 397:1042–1054, 2010.

[167] Alvin W. Hung, H. Leonardo Silvestre, Shijun Wen, Guillaume P C George, Jennifer Boland, Tom L. Blundell, Alessio Ciulli, and Chris Abell. Optimization of Inhibitors of Mycobacterium tuberculosis Pantothenate Synthetase Based on Group Efficiency Analysis. *ChemMedChem*, 11:38–42, 2016.

[168] Alicia P. Higueruelo, Adrian Schreyer, G Richard J Bickerton, Tom L. Blundell, and Will R. Pitt. What can we learn from the evolution of protein-ligand interactions to aid the design of new therapeutics? *PLoS One*, 7:e51742, 2012.

[169] Tjelvar S. G. Olsson, Mark A. Williams, Will R. Pitt, and John E. Ladbury. The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *J. Mol. Biol.*, 384:1002–17, 2008.

[170] LLC Schrödinger. The PyMOL Molecular Graphics System, Version 1.8. 2015.

[171] Daniel A. Erlanson. Calculating Hotspots in Detail, 2016.

[172] Jaconb Otto, Mark; Thornton. Bootstrap.

[173] Alexander S. Rose and Peter W. Hildebrand. NGL Viewer: A web application for molecular visualization. *Nucleic Acids Res.*, 43:W576–W579, 2015.

[174] Stefan Schmitt, Manfred Hendlich, and Gerhard Klebe. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chemie - Int. Ed.*, 40:3141–3144, 2001.

[175] Timo Krotzky and Gerhard Klebe. Acceleration of Binding Site Comparisons by Graph Partitioning. *Mol. Inform.*, 34:550–558, 2015.

[176] Timo Krotzky, Christian Grunwald, Ute Egerland, and Gerhard Klebe. Large-scale mining for similar protein binding pockets: With RAPMAD retrieval on the fly becomes real. *J. Chem. Inf. Model.*, 55:165–179, 2015.

[177] Chia-en a Chang, Wei Chen, and Michael K. Gilson. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U. S. A.*, 104:1534–1539, 2007.

[178] Stefan G. Krimmer, Jonathan Cramer, Michael Betz, Veronica Fridh, Robert Karlsson, Andreas Heine, and Gerhard Klebe. Rational Design of Thermodynamic and Kinetic Binding Profiles by Optimizing Surface Water Networks Coating Protein-Bound Ligands. *J. Med. Chem.*, 59:10530–10548, 2016.

[179] Jonathan Cramer, Stefan G. Krimmer, Andreas Heine, and Gerhard Klebe. Paying the Price of Desolvation in Solvent-Exposed Protein Pockets: Impact of Distal Solubilizing Groups on Affinity and Binding Thermodynamics in a Series of Thermolysin Inhibitors. *J. Med. Chem.*, 60:acs.jmedchem.7b00490, 2017.

[180] Tjelvar S. G. Olsson, John E. Ladbury, Will R. Pitt, and Mark A. Williams. Extent of enthalpy-entropy compensation in protein-ligand interactions. *Protein Sci.*, 20:1607–1618, 2011.

[181] Matteo Aldeghi, Alexander Heifetz, Michael J. Bodkin, Stefan Knapp, and Philip C. Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.*, 7:207–218, 2016.

[182] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, 2:3204–18, 2004.

[183] Jocelyn Sunseri and David Ryan Koes. Pharmit: interactive exploration of chemical space. *Nucleic Acids Res.*, 44:gkw287–, 2016.

[184] Ilenia Giangreco, Tjelvar S. G. Olsson, Jason C. Cole, and Martin J. Packer. Assessment of a Cambridge structural database-driven overlay program. *J. Chem. Inf. Model.*, 54:3091–3098, 2014.

[185] G Jones, P Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–48, 1997.

[186] Steven L. Dixon, Alexander M. Smondyrev, Eric H. Knoll, Shashidhar N. Rao, David E. Shaw, and Richard A. Friesner. PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput. Aided. Mol. Des.*, 20:647–671, 2006.

[187] P. C D Hawkins, A. Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, 50:74–82, 2007.

[188] Robert P. Sheridan and Simon K. Kearsley. Why do we need so many chemical similarity search methods? *Drug Discov. Today*, 7:903–911, 2002.

[189] Robert B. Murphy, Matthew P. Repasky, Jeremy R. Greenwood, Ivan Tubert-Brohman, Steven Jerome, Ramakrishna Annabhimoju, Nicholas A. Boyles, Christopher D. Schmitz, Robert Abel, Ramy Farid, and Richard A. Friesner. WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand-Receptor Docking. *J. Med. Chem.*, 59:4364–4384, 2016.

[190] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[191] Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Genet.*, 52:609–623, 2003.

[192] Robin Taylor, Jason C. Cole, Oliver Korb, and Patrick McCabe. Knowledge-based libraries for predicting the geometric preferences of druglike molecules. *J. Chem. Inf. Model.*, 54:2500–2514, 2014.

[193] Robin Taylor, Jason C. Cole, David A. Cosgrove, Eleanor J. Gardiner, Valerie J. Gillet, and Oliver Korb. Development and validation of an improved algorithm for overlaying flexible molecules. *J. Comput. Aided. Mol. Des.*, 26:451–472, 2012.

[194] Ajay N. Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *J. Comput. Aided. Mol. Des.*, 22:133–139, 2008.

[195] Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55:6582–6594, 2012.

[196] John J. Irwin and Brian K. Shoichet. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45:177–182, 2005.

[197] Oliver Korb, Thomas Stützle, and Thomas E. Exner. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. *J. Chem. Inf. Model.*, 49:84–96, 2009.

[198] John W. Liebeschuetz, Jason C. Cole, and Oliver Korb. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput. Aided. Mol. Des.*, 26:737–748, 2012.

[199] Kerim Babaoglu and Brian K. Shoichet. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.*, 2:720–3, 2006.

[200] Dima Kozakov, David R. Hall, Stefan Jehle, Lingqi Luo, Stefan O. Ochiana, Elizabeth V. Jones, Michael Pollastri, Karen N. Allen, Adrian Whitty, and Sandor Vajda. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proc. Natl. Acad. Sci.*, 112:E2585–E2594, 2015.

[201] Albert P. Li, Donald L. Kaminski, and Asenath Rasmussen. Substrates of human hepatic cytochrome P450 3A4. *Toxicology*, 104:1–8, 1995.

[202] B. Wu, E. Y. T. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov, and R. C. Stevens. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science (80-.).*, 330:1066–1071, 2010.

[203] Gebhard Thoma, Markus B. Streiff, Jiri Kovarik, Fraser Glickman, Trixie Wagner, Christian Beerli, and Hans Günter Zerwes. Orally bioavailable isothioureas block function of the chemokine receptor CXCR4 in vitro and in vivo. *J. Med. Chem.*, 51:7915–7920, 2008.

[204] Gerard J.P. van Westen, Anna Gaulton, and John P. Overington. Chemical, Target, and Bioactive Properties of Allosteric Modulation. *PLoS Comput. Biol.*, 10:e1003559, 2014.

[205] Jérémy Desaphy, Karima Azdimousa, Esther Kellenberger, and Didier Rognan. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.*, 52:2287–2299, 2012.

[206] Joseph C. Somody, Stephen S. MacKinnon, and Andreas Windemuth. Structural Coverage of the Proteome for Pharmaceutical Applications. *Drug Discov. Today*, 2017.

[207] Bernardo Ochoa-Montaño, Nishita Mohan, and Tom L. Blundell. CHOPIN: a web resource for the structural and functional proteome of Mycobacterium tuberculosis. *Database (Oxford).*, 2015:bav026–bav026, 2015.

[208] William S.J. Valdar. Scoring residue conservation. *Proteins Struct. Funct. Genet.*, 48:227–241, 2002.

[209] John A. Capra, Roman A. Laskowski, Janet M. Thornton, Mona Singh, and Thomas A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, 5:e1000585, 2009.

[210] Philip Jones, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex L. Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30:1236–1240, 2014.

[211] W Yu, Sirish K. Lakkaraju, E P Raman, and A D MacKerell Jr. Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling. *J Comput Aided Mol Des*, 28:491–507, 2014.

[212] Garrett M. Morris, Huey Ruth, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30:2785–2791, 2009.

[213] György G. Ferenczy and György M. Keseru. Thermodynamics of fragment binding. *J. Chem. Inf. Model.*, 52:1039–1045, 2012.

[214] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[215] S. Roy Kimura, Hai Peng Hu, Anatoly M. Ruvinsky, Woody Sherman, and Angelo D. Favia. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model.*, 57:1388–1401, 2017.

[216] Yusuke Kamada, Nozomu Sakai, Satoshi Sogabe, Koh Ida, Hideyuki Oki, Kotaro Sakamoto, Weston Lane, Gyorgy Snell, Motoo Iida, Yasuhiro Imaeda, Junichi Sakamoto, and Junji Matsui. Discovery of a B-Cell Lymphoma 6 Protein-Protein Interaction Inhibitor by a Biophysics-Driven Fragment-Based Approach. *J. Med. Chem.*, 60:4358–4368, 2017.